MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: ICER AND THE FAILURE OF MEASUREMENT IN REFERENCE CASE HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 4 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

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

The objective of this study is to evaluate whether the Institute for Clinical and Economic Review (ICER), as the most influential non-governmental health technology assessment body in the United States, satisfies the minimum requirements of scientific measurement and falsifiable evidence in its evaluation of therapy value. Using a 24-item diagnostic grounded in representational measurement theory, the study interrogates ICER's implicit belief structure rather than its stated methodological rhetoric. The aim is not to assess the elegance of ICER's models or the transparency of its processes, but to determine whether the quantitative claims ICER advances are admissible as measurements, capable of supporting arithmetic, aggregation, and empirical refutation. In particular, the analysis asks whether ICER's reliance on utilities, QALYs, and reference-case simulation reflects a coherent measurement framework or a systematic inversion in which arithmetic is treated as authoritative in the absence of valid measurement.

The findings are unequivocal and severe. ICER's belief profile exhibits one of the most extreme inversions of representational measurement observed across the HTA ecosystem. Core axioms including measurement preceding arithmetic, the necessity of ratio scales for multiplication, unidimensionality, and the inadmissibility of composite constructs are weakly endorsed or rejected outright, while mathematically impossible propositions embedded in ICER's reference-case framework are endorsed at or near the ceiling of the logit scale. Rasch measurement, the only defensible basis for latent-trait claims derived from patient-reported outcomes, is effectively absent. The resulting structure is not one of partial misunderstanding or technical dispute, but of categorical reversal: ICER treats arithmetic outputs from simulation models as evidentiary facts while systematically excluding the measurement conditions that would make those outputs scientifically meaningful. ICER does not merely participate in the HTA memeplex; it operationalizes it as pricing authority.

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 ICER KNOWLEDGE BASE

For the purposes of this analysis, the ICER knowledge base is defined as the recurring set of concepts, assumptions, modeling conventions, and evaluative norms that shape how ICER constructs, interprets, and defends claims about therapy value. It is not defined by individual reports or methodological appendices, but by the stable architecture that appears across disease areas, clinical contexts, and successive assessments. This architecture is reproduced through ICER's reference-case framework, its published value assessment reports, its engagement with payers and manufacturers, and its influence on U.S. academic and policy discourse.

At the core of this knowledge base is the routine acceptance of cost-utility analysis as a legitimate quantitative foundation for decision making. ICER treats utilities derived from preference-based instruments as if they were interval or ratio measures, suitable for multiplication by time and aggregation across individuals. QALYs are assumed to be dimensionally homogeneous quantities, capable of lawful arithmetic and population-level aggregation. These assumptions are never treated as hypotheses requiring validation; they function as axioms internal to the ICER framework. Measurement theory is not invoked as a constraint on model construction, but is implicitly displaced by convention and precedent.

Reference-case simulation modeling occupies a central epistemic role within the ICER knowledge base. Models are treated as engines of evidence rather than as exploratory devices. Sensitivity analyses are presented as demonstrations of robustness rather than as illustrations of assumption dependence. The distinction between conditional projection and empirical claim is systematically blurred. As a result, model outputs acquire normative authority despite being insulated from falsification. Claims about long-term cost-effectiveness, value thresholds, and pricing benchmarks are advanced without any requirement that the underlying quantities be measurable in the scientific sense.

Patient-reported outcomes are incorporated into this framework in a similarly non-measurement-based manner. Ordinal questionnaire responses are summed, indexed, or mapped through preference algorithms and then treated as quantitative inputs. The ICER knowledge base exhibits no engagement with Rasch measurement or conjoint simultaneous measurement, despite heavy reliance on latent constructs such as quality of life, symptom burden, and patient experience. The transformation from ordinal response to arithmetic quantity is accomplished by algorithmic convention rather than by satisfaction of measurement axioms. Subjective data are therefore monetized without ever being measured.

Equally important are the silences that define the ICER knowledge base. Representational measurement theory is absent from ICER's methodological vocabulary. Scale-type constraints, invariance requirements, and the logical conditions for aggregation are not discussed as limiting

principles. Falsification is invoked rhetorically but redefined operationally to mean stability across model scenarios rather than exposure to empirical refutation. The evolution of objective knowledge is replaced by iterative model refinement.

Taken together, these elements define a coherent but scientifically inadmissible knowledge base. ICER's authority rests not on measurement, but on the disciplined repetition of a shared belief system that treats arithmetic as self-justifying. The 24-item diagnostic shows that this structure is not accidental or transitional. It is stable, internally reinforced, and actively transmitted to payers, policymakers, and the broader HTA community.

.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: INSTITUTE FOR CLINICAL AND ECONOMIC REVIEW

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 INSTITUTE FOR CLINICAL AND ECONOMIC REVIEW

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

MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.90	+2.20
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.95	+2.50
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.95	+2.50
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	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.95	+2.50
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.90	+2.20
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	0.95	+2.50
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	0.60	+0.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.20	-1.40
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

ICER: ARITHMETIC WITHOUT MEASUREMENT AS INSTITUTIONAL PRACTICE

The purpose of this assessment is not to argue that the Institute for Clinical and Economic Review (ICER) occasionally misapplies methods or adopts contestable modeling assumptions. That would trivialize what the 24-item diagnostic reveals. The objective is to determine whether ICER's analytic framework satisfies the axioms required for scientific measurement, lawful arithmetic, falsification, and the cumulative evolution of objective knowledge. On this criterion, the results are unambiguous. ICER does not merely fall short of representational measurement standards; it has constructed an evaluative architecture that systematically inverts them. Measurement is treated as optional, arithmetic as authoritative, and simulation as evidence. The resulting logit profile does not describe methodological disagreement. It describes epistemic collapse.

The defining feature of the ICER profile is the categorical rejection of measurement as a prerequisite for calculation (Table 1). The proposition that measurement must precede arithmetic is endorsed at $p = 0.10$, corresponding to a canonical logit of -2.20. This is not a marginal deviation. It places ICER at the extreme rejection boundary of a principle that has governed quantitative science since the emergence of representational measurement theory. Once this rejection is made, everything that follows is mechanically determined. Arithmetic no longer requires justification. Numbers become legitimate by virtue of being computable, not by virtue of representing quantities.

This inversion explains the near-ceiling endorsement of propositions that are mathematically impossible under any coherent theory of measurement. ICER endorses, at the maximum or near-maximum logit values, the beliefs that utilities are ratio measures, that ratio measures can take negative values, that EQ-5D algorithms generate interval measures, that summated Likert responses create ratio quantities, and that QALYs can be aggregated across individuals and populations. These propositions are not peripheral. They are the load-bearing assumptions of ICER's reference-case simulation model. Without them, the ICER framework collapses immediately. Incremental cost-effectiveness ratios become undefined, thresholds lose coherence, and "value-based price benchmarks" dissolve into arithmetic fiction.

The diagnostic shows that ICER resolves this threat not by defending these propositions, but by excluding the axioms that would invalidate them. The proposition that multiplication requires a ratio measure sits at p = 0.10 (−2.20). The proposition that meeting representational axioms is required for arithmetic sits at the same value. These are not subtle errors. They represent an explicit institutional commitment to performing arithmetic on non-measures. ICER's calculations are therefore not just questionable; they are meaningless by construction.

The treatment of unidimensionality illustrates how this meaninglessness is normalized. Measures must be unidimensional is endorsed at p = 0.15 (−1.75). Yet ICER routinely treats composite health-state descriptions, derived from multiple heterogeneous domains, as if they represented a single quantity. Health-related quality of life is treated as a scalar attribute despite being assembled from mobility, pain, anxiety, self-care, and other qualitatively distinct dimensions. This is not measurement. It is category error. The diagnostic shows that ICER resolves this by denying the requirement rather than confronting the violation.

Nowhere is this clearer than in ICER's treatment of latent traits and patient-reported outcomes. The Rasch block collapses entirely to the floor of the scale. Every proposition that would impose measurement discipline on subjective claims is endorsed at p = 0.05, with canonical logits of −2.50. ICER rejects, categorically, the proposition that there are only two admissible quantitative measures: linear ratio scales for manifest attributes and Rasch logit ratio scales for latent traits. It rejects the proposition that transforming subjective responses to interval measurement is only possible with Rasch rules. It rejects the proposition that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits. These values indicate structural exclusion, not methodological debate.

This exclusion is decisive. Rasch measurement is not one option among many. It is the only known transformation model capable of producing invariant measurement from ordinal subjective responses. By rejecting Rasch while endorsing summation-based scoring, ICER institutionalizes pseudo-measurement. Subjective responses are not measured; they are scored, averaged, multiplied, and monetized without ever satisfying the conditions of measurement. Patient experience is invoked rhetorically while being numerically abused.

The outcome of interest for latent traits makes this explicit. The proposition that the outcome of interest is possession of the trait is endorsed at only p = 0.20 (−1.40). ICER prefers to speak in terms of changes in scores, differences in means, and modeled quality-adjusted life expectancy rather than confronting the substantive question of what it means to possess more or less of a latent attribute. This avoidance is not accidental. Possession requires invariance. Invariance requires Rasch. This framework would dismantle ICER's scoring conventions. The system therefore rejects possession conceptually in order to preserve arithmetic operationally.

Aggregation completes the epistemic failure. ICER endorses, at the maximum logit value of +2.50, the proposition that QALYs can be aggregated. Aggregation is not a technical convenience; it is the mechanism by which ICER converts individual-level pseudo-quantities into population-level pricing authority. Yet aggregation requires dimensional homogeneity and ratio-scale properties that ICER explicitly denies are required. This is not inconsistency. It is a belief system in which arithmetic outcomes are privileged over measurement constraints.

The reference-case simulation model is the institutional mechanism through which this belief system is operationalized. ICER endorses, at $p = 0.95$ (+2.50), the belief that reference-case simulations generate falsifiable claims. They do not. Simulation outputs are conditional projections derived from assumptions, many of which rest on non-measured inputs. Sensitivity analysis does not confer falsifiability. It explores internal model behavior. A claim that cannot be wrong in the empirical sense is not a scientific claim. ICER's endorsement of simulation as falsifiable evidence represents a fundamental misunderstanding of scientific risk.

This misunderstanding allows ICER to occupy a unique and dangerous position in U.S. health policy. Although it lacks statutory authority, its outputs function as de facto price controls. Therapies are labeled "low value" or "high value" based on modeled ICERs derived from non-measures. Payers, health systems, and policymakers defer to these labels not because they are empirically defensible, but because they are numerically confident. Arithmetic substitutes for evidence. Authority substitutes for truth.

The most damning aspect of the ICER profile is not ignorance. It is selectivity. ICER exhibits perfect clarity when dealing with physical quantities. Time is recognized, correctly and emphatically, as a ratio measure, endorsed at $p = 0.95$ (+2.50). No confusion exists there. The confusion appears only when arithmetic threatens to invalidate preferred constructs. This selectivity demonstrates that ICER is not incapable of understanding measurement. It is unwilling to apply it where it would disrupt institutional practice.

This unwillingness has predictable consequences for the evolution of knowledge. Without measurement, there can be no cumulative science. Claims cannot be replicated in the strong sense because there is no invariant quantity to reproduce. Disagreements are resolved through negotiation, scenario analysis, or appeals to "reasonableness" rather than empirical refutation. What evolves is not objective knowledge, but consensus around a shared fiction. ICER's analytic output is therefore not evidence in the scientific sense. It is narrative stabilized by arithmetic.

Defenders of ICER often invoke pragmatism. Decision makers "need numbers," and ICER supplies them. This argument collapses immediately. Measurement axioms do not prevent decision making. They prevent inadmissable application of numbers. A system that rejects these axioms does not become practical; it becomes unaccountable. ICER's numbers cannot be wrong because they are not grounded in measurable quantities. They can only be replaced by different numbers generated under different assumptions.

The conclusion forced by this diagnostic is severe but unavoidable. ICER does not conduct health technology assessment in any scientifically defensible sense. It conducts numerical storytelling, built on arithmetic without measurement, simulation without falsification, and aggregation without dimensional coherence. Its reference-case model is not a tool for evidence synthesis. It is an engine for producing authoritative-looking ratios that cannot, in principle, represent anything real.

If ICER were to accept the measurement critique, its entire framework would have to be dismantled. QALYs would be abandoned. Utilities would be reclassified as ordinal preferences. Reference-case simulations would also be abandoned. Manifest claims would be restricted to linear

ratio measures. Latent traits would require Rasch logit ratio measurement with demonstrated invariance. Price benchmarks would disappear. That is why reform has not occurred.

The 24-item logit profile shows that ICER's failure is not accidental, remediable, or transitional. It is structural. ICER apparently exists to perform arithmetic on non-measures and to present the results as evidence. In representational measurement terms, that is not a methodological weakness. It is total failure.

DOES ICER HAVE A LEGACY? ARE THERE LESSONS TO BE LEARNED FOR HEALTH TECHNOLOGY ASSESSMENT?

ICER unquestionably has a legacy, but it is not the legacy its founders or defenders would claim. It is not a legacy of methodological rigor, scientific advance, or disciplined evaluation. It is a legacy of institutionalizing arithmetic without measurement and demonstrating how numerical storytelling can be converted into durable policy influence. ICER's significance lies not in the validity of its outputs, but in the fact that it shows how far a health system can travel from the axioms of science while still persuading itself that it is acting "evidence based."

The most important lesson from ICER is that technical sophistication can substitute for scientific legitimacy when the audience lacks measurement literacy. ICER did not invent cost-effectiveness analysis, QALYs, or reference-case simulation. Those tools were already present in the HTA memeplex. ICER's contribution was to consolidate them into a single, highly standardized evaluative template and then present the resulting outputs as authoritative pricing signals. By doing so, it revealed a structural vulnerability in HTA: when representational measurement theory is absent from professional training, almost any internally coherent model can be mistaken for evidence.

ICER also demonstrates how an institution can gain power without formal authority. It has no statutory mandate, yet its reports shape pricing negotiations, formulary placement, and access decisions across the United States. This influence rests entirely on the perception that ICER's numbers are grounded in objective measurement. The diagnostic evidence shows that this perception is false. The lesson is not merely that ICER is wrong, but that health systems will defer to numerical outputs even when the numbers lack the properties required to support arithmetic, aggregation, or falsification. Authority migrates to whoever supplies numbers that look scientific.

Another lesson is that methodological consensus can form without scientific testing. ICER's reference-case framework is defended through repetition, citation, and alignment with international practice, not through empirical validation. The fact that ICER's core constructs violate elementary measurement axioms has never triggered an internal crisis or substantive reform. Instead, criticism is deflected as philosophical, impractical, or irrelevant. This illustrates how a memeplex stabilizes itself: challenges that threaten its foundational assumptions are excluded rather than engaged. The absence of debate is not evidence of correctness; it is evidence of closure.

ICER's legacy also exposes the dangers of conflating policy usefulness with scientific validity. Defenders often argue that ICER's analyses are "useful for decision making," even if they are imperfect. This argument collapses the distinction between acting and knowing. Decisions can always be made; the question is whether the claims informing them are true, falsifiable, and capable of supporting cumulative knowledge. ICER's framework produces decisions, but it does not produce knowledge. Its simulations cannot be wrong in the scientific sense because they are insulated from empirical refutation. The lesson for HTA is stark: usefulness without measurement is not pragmatism, it is epistemic surrender.

Perhaps the most sobering lesson is how difficult it is to reverse such a legacy once it is entrenched. ICER's methods are now embedded in journals, curricula, payer expectations, and international dialogue. Entire professional identities have formed around fluency in ICER-style modeling. Acknowledging that the foundation is unsound would invalidate decades of work and authority. Institutions rarely choose self-invalidating reform. This explains why measurement axioms are rejected at the belief level rather than addressed analytically. The cost of correction is too high.

Yet ICER's legacy also provides a warning and an opportunity. It shows that HTA has reached a point where further refinement of models, thresholds, and scenarios adds nothing of scientific value. The path forward cannot be incremental. Either HTA recommits to representational measurement, falsifiability, and the evolution of objective knowledge, or it accepts that it is a policy craft rather than a science. ICER forces that choice into the open.

In that sense, ICER's enduring contribution may be negative but clarifying. It demonstrates the endpoint of a discipline that elevates arithmetic over measurement and consensus over truth. The lesson for HTA is not how to build better reference-case models, but why such models must be abandoned as evidentiary devices. If HTA is to have a future as a scientific enterprise, ICER's legacy must be understood not as a model to emulate, but as a cautionary tale of what happens when measurement is treated as optional.

?

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