MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

IRELAND: THE UNCRITICAL OPERATIONAL INHERITANCE OF NICE FALSE MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 249 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The National Centre for Pharmacoeconomics (NCPE) is Ireland's national health technology assessment authority responsible for evaluating the clinical and economic value of pharmaceutical therapies to inform reimbursement and pricing decisions within the publicly funded health system. Established in 1998 and based at St. James's Hospital in Dublin, the NCPE operates under the auspices of the Health Service Executive (HSE) and provides independent assessments to support decisions on whether new medicines should be publicly funded under Ireland's community drug schemes and hospital reimbursement programs.

The NCPE's primary function is to conduct pharmacoeconomic evaluations, including cost-effectiveness and budget impact analyses, submitted by pharmaceutical manufacturers seeking reimbursement. These evaluations typically employ incremental cost-effectiveness ratios, often expressed as cost per quality-adjusted life year (QALY), to compare new therapies with existing standards of care. The Centre also provides rapid reviews, methodological guidance to manufacturers, and formal health technology assessments for high-cost or high-impact therapies.

Through this process, the NCPE plays a decisive operational role in determining patient access to medicines in Ireland. Its assessments inform HSE reimbursement negotiations, pricing agreements, and final funding decisions. As such, the NCPE functions as the principal quantitative gatekeeper for pharmaceutical adoption within Ireland's national health system, translating economic evaluation methods into binding policy recommendations affecting therapy availability and resource allocation.

The objective of this study was to determine whether the National Centre for Pharmacoeconomics (NCPE), as Ireland's national HTA authority responsible for evaluating pharmaceutical therapies and informing reimbursement decisions, applies quantitative constructs that satisfy the axioms of representational measurement. The NCPE occupies a decisive operational role in the Irish health system. Its methodological guidance determines the structure of economic submissions, the preferred outcome measures for therapy evaluation, and the quantitative basis upon which reimbursement recommendations are made to the Health Service Executive. This analysis applied the 24-item canonical representational measurement diagnostic to the NCPE knowledge base, including its pharmacoeconomic guidelines, submission requirements, and evaluation framework. The objective was not to evaluate procedural rigor or administrative transparency, but to determine whether the NCPE enforces the foundational conditions required for lawful arithmetic, including unidimensionality, invariant unit structure, dimensional homogeneity, and the requirement that multiplication and division operate only on ratio measures. The study therefore addresses a decisive question: whether Ireland's operational HTA authority evaluates therapies using measurement-valid constructs capable of supporting empirically meaningful quantitative claims.

The logit profile demonstrates systematic exclusion of representational measurement axioms within the operational knowledge base of the NCPE. Core statements asserting that measurement must precede arithmetic, that multiplication requires ratio scale properties, that latent constructs require Rasch transformation to establish invariant interval structure, and that cost-effectiveness claims based on composite indices fail representational measurement requirements collapse to floor or near-floor logit values, indicating effective non-possession. Conversely, false statements asserting that QALYs constitute ratio measures, that composite utility scores may be aggregated and manipulated arithmetically, and that simulation-derived cost-effectiveness ratios represent empirically evaluable quantities register positive logit values, indicating operational endorsement. These findings establish that the NCPE's evaluative architecture is structurally dependent on quantitative constructs that do not satisfy the axioms required for empirical measurement. Arithmetic operations are performed on composite ordinal scores derived from multiattribute instruments, and simulation outputs are treated as evaluative evidence despite lacking empirical falsifiability. The resulting framework constitutes a coherent administrative system but lacks the defining properties of measurement-based scientific evaluation; focused instead on numerical storytelling.

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 NATIONAL CENTRE FOR PHARMACOECONOMICS KNOWLEDGE BASE

The knowledge base of the National Centre for Pharmacoeconomics is defined by its role as Ireland’s operational authority for pharmacoeconomic evaluation and reimbursement assessment. The Centre provides methodological guidance to pharmaceutical manufacturers submitting therapies for public reimbursement and conducts independent evaluations to inform funding decisions by the Health Service Executive. Its framework is designed to support comparative assessment of therapies across disease areas and clinical indications, integrating clinical effectiveness evidence with economic evaluation to generate quantitative claims regarding therapeutic value.

At the core of the NCPE’s evaluative framework is the requirement that manufacturers submit incremental cost-effectiveness analyses, typically expressed as cost per quality-adjusted life year (QALY) gained. These analyses combine estimates of incremental costs and incremental health outcomes, derived from clinical trials, observational studies, and simulation modeling. Health outcomes are quantified using utility weights obtained from preference-based multiattribute instruments, most commonly the EQ-5D. These utility weights are applied to survival time to generate QALYs, which function as summary measures of therapy impact. Incremental cost-effectiveness ratios derived from these calculations serve as central evaluative metrics in reimbursement decision-making.

Simulation modeling plays a central role in operationalizing these constructs. Models integrate clinical evidence, epidemiological assumptions, treatment pathways, and utility weights to project long-term outcomes beyond the duration of observed clinical trials. These projections generate estimates of lifetime costs and QALYs, allowing comparison of therapies in terms of incremental cost-effectiveness. The resulting quantitative outputs inform reimbursement recommendations and pricing negotiations between the Health Service Executive and pharmaceutical manufacturers.

The NCPE framework presents itself as a structured and transparent system designed to ensure rational allocation of healthcare resources. Its methodological guidance specifies submission requirements, modeling standards, and reporting conventions, providing consistency across evaluations. This architecture reflects methodological alignment with international HTA practice, particularly frameworks developed in the United Kingdom and adopted across European health systems.

However, the quantitative constructs embedded within this framework originate from composite utility indices derived from ordinal preference elicitation over multidimensional health states. These utility scores represent aggregated preferences rather than invariant measures of a unidimensional attribute. Their numerical properties are determined by scoring algorithms and valuation conventions rather than demonstration of measurement invariance or ratio scale structure. Despite this, the NCPE framework permits multiplication of utility scores by time and

aggregation across individuals, treating the resulting values as if they possessed ratio scale properties.

The knowledge base therefore reflects a coherent administrative structure built upon inherited quantitative conventions. It demonstrates procedural rigor in applying established economic evaluation methods but does not require demonstration that the constructs used satisfy representational measurement axioms. Composite utility scores function as operational decision variables despite lacking invariant unit structure. Simulation outputs serve as evaluative evidence despite lacking empirical falsifiability.

As a result, the NCPE knowledge base embodies an evaluative framework in which quantitative sophistication and administrative coherence coexist with structural absence of measurement foundations. Numerical outputs acquire operational authority within reimbursement decision-making, yet their measurement properties remain assumed rather than demonstrated. NCPE has inherited a playbook for numerical storytelling.

.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: IRELAND: NATIONAL CENTRE FOR PHARMACOECONOMICS

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.

NUMERICAL STORYTELLING AND THE IRISH NATIONAL CENTRE FOR PHARMACOECONOMICS

The canonical statement logit profile of Ireland's National Centre for Pharmacoeconomics reveals not an incomplete understanding of measurement principles but their systematic exclusion as operational constraints within the evaluative framework used to determine pharmaceutical access and reimbursement (Table 1). The pattern is structurally familiar because the NCPE functions within a methodological architecture that closely tracks the NICE reference case. The importance of this finding is not that Ireland has "copied" the UK, which it has, but that copying here means adopting the same inversion of logic: arithmetic is performed first, and measurement, if acknowledged at all is treated as a descriptive afterthought rather than the necessary precondition for quantitative claims. In any quantitative science, measurement precedes arithmetic. In the NCPE knowledge base, the logit profile shows that this order is reversed to support therapy impact claims as numerical storytelling..

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS IRELAND NATIONAL CENTRE FOR PHARMACOECONOMICS

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	0.90	+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	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.05	-2.50
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	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.95	+2.50

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.65	+0.60
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.55	+0.50
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.55	+0.50
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.10
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The repeated floor collapses are decisive for NCPE. Statements asserting that meeting the axioms of representational measurement is required for arithmetic, that there are only two defensible classes of measurement (linear ratio for manifest attributes and Rasch logit ratio for latent traits), and that transforming subjective responses to interval measurement is only possible under Rasch rules all register at −2.50. In the logic of your diagnostic, −2.50 denotes effective non-possession: the proposition does not operate as a binding principle within the knowledge base, irrespective of whether it might appear occasionally in narrative form. These are not marginal propositions. They are the structural conditions under which numbers can legitimately represent quantities. Their collapse to the floor therefore indicates that the NCPE evaluative framework does not treat measurement axioms as constraints on what it may calculate, report, or use in decision-making.

This non-possession is paired with strong positive endorsement of the very propositions required to sustain the simulated reference-case apparatus. The QALY-as-ratio claim registers at +2.50, QALY aggregation at +2.50, and the claim that reference case simulations generate falsifiable claims at +2.20. These positive logits indicate that the NCPE knowledge base treats these propositions not as contingent assumptions but as normal, legitimate, and operationally admissible. The structure is not ambiguous. The agency-level knowledge base simultaneously excludes the axioms that constrain arithmetic while endorsing the constructs that require those axioms to be ignored. That is the definition of institutionalization. The NCPE does not merely tolerate false measurement; it operationalizes it as the mischaracterized quantitative foundation for reimbursement evaluation.

The measurement failure is not subtle. It is built into the definition of the central outcome metric. A QALY combines time, correctly recognized in the table as a manifest ratio measure, with a utility weight derived from preference elicitation over multidimensional health state descriptions.

Those preference weights are not established as measures of a unidimensional attribute. They are scoring conventions derived from valuation surveys and algorithmic transformations of responses to multiattribute descriptive systems. Even if one were to grant interval properties to these weights, an unjustified concession without Rasch transformation, interval scales do not support multiplication. Multiplication requires ratio scales with a true zero. The NCPE knowledge base endorses multiplication nonetheless. The logit profile demonstrates that the statement "multiplication requires a ratio measure" is recognized as true in principle but does not function as an operational constraint, collapsing to −2.20. The result is a system that treats scoring outputs as if they were quantities and then performs arithmetic operations that are inadmissible under the scale properties of those outputs; a requirement formalized by Stevens in 1946.

Dimensional homogeneity further exposes the incoherence. The QALY is not a dimensionally homogeneous quantity because it is produced by multiplying a ratio measure (time) by a composite preference score that lacks demonstrated ratio properties and is not a measure of a single attribute. A "year" is a measurable quantity. A preference score is not a measurable quantity unless transformed into an invariant unit structure on a unidimensional continuum. Combining them produces a composite mathematically meaningless number. Yet the NCPE knowledge base endorses the proposition that the QALY is dimensionally homogeneous (+2.20). This is not a minor technical error. Dimensional homogeneity is the firewall that prevents arithmetic from combining unlike quantities and calling the result a measure. The positive logit here indicates that the firewall is not merely absent; it is inverted into its opposite. The system treats the composite as if it were a coherent quantity and then proceeds to compute ratios, thresholds, and policy conclusions.

The resource side of the cost-effectiveness apparatus adds a second layer of quantitative disorder. Monetary "cost" is commonly treated as a ratio measure, but its use in cost-effectiveness modeling typically bundles heterogeneous resource inputs into a composite monetary total. While money itself has ratio properties, it is not invariant across settings, and it functions as a price-weighted composite rather than as a measured resource unit. If the objective is to support replicable, falsifiable claims about health system impact, the appropriate basis is not composite cost totals but resource units (hospital days, admissions, procedures, physician visits, device utilization), each of which can be specified as a linear ratio measure and empirically tracked within defined target populations. The NCPE framework, however, privileges the integrative fiction of a single cost-per-QALY ratio, which requires both sides of the ratio to be treated as homogeneous quantities. The logit profile indicates that the knowledge base does not recognize the necessity of such structural discipline. NCPE instead treats the resulting ratio as a legitimate object of decision-making.

The asymmetry in the profile is instructive. Time is properly recognized as a ratio measure (+2.50). This shows that the knowledge base is capable of recognizing lawful measurement structures for manifest attributes. The failure is therefore not necessarily ignorance of measurement per se. It is selective suspension of measurement constraints precisely where the evaluative framework requires that suspension in order to function, namely, in the treatment of multiattribute constructs such as "quality of life" and in the aggregation of multiattribute preferences into a single score. This selective suspension is exactly what defines the HTA memeplex: a closed evaluative system that must reject binding measurement principles in order to preserve its core arithmetic operations.

The Rasch items sharpen this conclusion. If latent traits are to be quantified, if one is to claim to measure symptom burden, functioning, need fulfillment, or quality of life, then ordinal responses must be transformed into invariant measures. Rasch measurement is not a psychometric preference but the only defensible transformation model that yields invariant unit structure and separable item and person parameters. The repeated −2.50 floor values on Rasch-related statements demonstrate that the NCPE knowledge base does not treat this requirement as relevant. It has no notion of Rasch measurement. This matters because HTA routinely makes therapy impact claims in terms of patient-reported outcomes and preference-weighted constructs. Without Rasch transformation, these are not measures. They are scores. Yet scores are treated as if they were quantities, aggregated, multiplied, and inserted into thresholds. The logit profile therefore exposes a decisive structural contradiction: the NCPE framework claims to quantify therapy impact using constructs that are not measured and then performs arithmetic that presupposes they are. The fact that Rasch modelling was first proposed in 1960 and shown to be equivalent in its rules to the axioms of representational measurement in 1977 is, apparently, of no consequence.

The falsifiability items reveal the epistemic consequence. The NCPE knowledge base shows positive endorsement for the statement that non-falsifiable claims should be rejected (+0.60). That looks, superficially, like a commitment to normal science. But it is immediately contradicted by the strong positive endorsement of the proposition that reference case simulations generate falsifiable claims (+2.20). Simulations do not generate falsifiable claims in the scientific sense. They generate outputs conditional on assumptions. They can be recalculated, re-parameterized, and extended indefinitely, but they cannot be proven wrong in the way that empirical claims about measured quantities can be proven wrong. When a model fails to match reality, the failure is absorbed as "uncertainty," "parameter revision," or "structural sensitivity." Error does not trigger rejection; it triggers refinement. This is the hallmark of closure. The logit profile shows that the NCPE evaluative system has substituted internal model coherence for empirical testability and then labeled that substitution "evidence." The scientific revolution of the 17th century with its emphasis on unidimensionality and hypothesis testing is just cast aside as irrelevant to NCPE decision making.

This rejection has direct implications for duty of care. Reimbursement decisions determine therapy access. They determine which patients will receive an intervention, under what restrictions, at what speed, and at what negotiated price. If the quantitative foundation of those decisions is built on constructs that do not satisfy measurement axioms, then the system's numerical authority is detached from empirical reality. That detachment is not ethically neutral. It means that patient access can be constrained by administrative numbers that lack quantitative meaning. It means that physicians can be asked to justify therapy choices within an evaluative regime that treats composite preference scores as if they were measures. And it means that health systems can be led to believe that they are allocating resources rationally when the quantitative comparisons on which those allocations rest are not grounded in admissible arithmetic. A system where numerical storytelling is the key decision variable. NCPE could surely do better.

The most important scientific casualty is the evolution of objective knowledge. In normal science, claims are made in measurable terms, subjected to replication, and exposed to the risk of being wrong. Measurement enables cumulative learning because it stabilizes units and permits meaningful comparison across time, settings, and populations. The NCPE knowledge base, by

contrast, anchors decision-making in constructs that cannot be measured and in simulation outputs that cannot be falsified. Under these conditions, there is no pathway to cumulative correction. There is only procedural reiteration. A model can be updated, but the foundational error, the absence of measurement validity, remains. This is why your logit floor values matter. They show that the measurement axioms required for correction do not function as constraints. Without constraints, there is no scientific learning. There is only administrative continuity; one numerical story establishing closure after another.

Ireland's case is therefore analytically valuable because it demonstrates how the HTA false measurement memeplex propagates. The NCPE is not a fringe institution; it is an operational gatekeeper for access to medicines within a modern European health system. Its adoption of NICE-derived constructs demonstrates that methodological diffusion in HTA is not diffusion of validated measurement practice but diffusion of an administratively useful but nonsensical scoring framework. The framework spreads because it provides closure: a single index, a single ratio, a single decision logic that can be applied across therapies. But closure is not science. Closure is the avoidance of falsification. It is the refusal to expose claims to the risk of being wrong.

What would a measurement-valid alternative look like? It would begin by disallowing arithmetic on non-measures. It would restrict evaluable claims to unidimensional attributes. For manifest outcomes, it would require linear ratio scales: survival time, admissions avoided, hospital days, procedures, clinician visits, medication possession, and other resource units that can be measured and replicated within defined target populations over defined timeframes. For latent outcomes, it would require Rasch logit ratio scales built from instruments designed to meet invariance and unidimensionality. Claims would be protocol-driven and empirically evaluable, not model-driven and assumption-dependent. Therapy impact would be assessed through a portfolio of single-claim protocols rather than a single integrative fiction.

The NCPE logit profile shows that this alternative architecture is excluded at the level that matters: as an operational constraint. The repeated floor values on Rasch and representational measurement axioms indicate that these requirements do not function as admissibility conditions. The system therefore cannot self-correct from within its own evaluative logic. It can only refine the same constructs and expand the same modeling conventions. In that sense, Ireland is not merely adopting NICE methodology; it is adopting NICE's epistemic closure. It is locked into validating one numerical story after another.

The conclusion is not that Ireland's HTA authority lacks professionalism or technical competence. The conclusion is that technical competence applied to an invalid measurement framework cannot rescue the validity of the outputs. Complexity does not generate measurement. Consensus does not generate ratio properties. Sensitivity analysis does not generate falsifiability. Only lawful measurement produces quantities. Only quantities support admissible arithmetic. The NCPE knowledge base, as revealed by the logit profile, inverts this order. It treats arithmetic as primary and measurement as optional. That is why the profile shows systematic non-possession of measurement axioms alongside strong endorsement of QALY arithmetic.

If NCPE oi to claim scientific legitimacy as a reputable center for therapy impact assessment, and to meet its duty of care to patients, clinicians, and health systems, it must adopt the same constraints

that govern every other quantitative field. The axioms of representational measurement are not negotiable. They are not “one approach among many.” They are the conditions under which numbers can represent empirical attributes. Until the NCPE and the broader European HTA system enforce these axioms as operational constraints, their quantitative outputs will continue to possess administrative authority without measurement validity; a library of numerical stories. A decision framework built on such outputs cannot claim to be a science of therapy assessment; it can only claim to be a numerically organized administrative practice that is outside the remit of normal science. An activity that makes no contribution whatsoever to the evolution of objective knowledge.

III. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

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

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

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

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

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

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

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

MEANINGFUL THERAPY IMPACT CLAIMS

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

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

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

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

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

THE PATH TO MEANINGFUL MEASUREMENT

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

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

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

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

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

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

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

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

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

TRANSITION REQUIRES TRAINING

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

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

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

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

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

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

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

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

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

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

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

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

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

DESIGNED FOR CLOSURE

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

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

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

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

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

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

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

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

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

ACKNOWLEDGEMENT

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

REFERENCES

[1] 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