MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

CZECH REPUBLIC: THE NATIONAL ACCEPTANCE OF 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 420 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

Health technology assessment (HTA) in the Czech Republic is embedded within a centralized statutory framework that governs pharmaceutical pricing, reimbursement, and market access. The operational responsibility rests primarily with the State Institute for Drug Control (SÚKL), which functions as the national regulatory and reimbursement authority for medicinal products. SÚKL evaluates clinical effectiveness, safety, and economic value to determine whether new therapies qualify for reimbursement under the country's public health insurance system. Its assessments directly influence both reimbursement status and the level of reimbursement granted.

Economic evaluation plays a formal and increasingly important role in this process. Manufacturers seeking reimbursement for innovative or high-cost therapies are required to submit pharmacoeconomic analyses, typically including cost-effectiveness and budget impact assessments. These evaluations often use incremental cost-effectiveness ratios, expressed in terms of cost per quality-adjusted life year (QALY), to support claims of therapeutic and economic value. The Czech HTA framework therefore aligns closely with broader European methodological standards, reflecting influence from agencies such as NICE and EUnetHTA.

Final reimbursement and pricing decisions are made within a statutory administrative process involving SÚKL and health insurance funds, operating under national pharmaceutical legislation. HTA in the Czech Republic thus serves as a formal gatekeeping mechanism, determining patient access to new therapies while supporting resource allocation decisions within a publicly funded health system.

The objective of this study was to evaluate whether the national health technology assessment (HTA) knowledge base in the Czech Republic recognizes and operationalizes the axioms of representational measurement as necessary preconditions for quantitative evaluation of therapy impact. Using the 24-item canonical representational measurement diagnostic, the analysis interrogated national-level HTA materials, including reimbursement guidance, pharmacoeconomic submission requirements, academic publications, and methodological frameworks embedded within Czech HTA practice. The purpose was not to evaluate administrative sophistication or procedural completeness, but to determine whether the Czech HTA knowledge environment recognizes the foundational distinction between measurement and scoring. Specifically, the study assessed whether core measurement requirements to include unidimensionality, invariant unit structure, dimensional homogeneity, and the requirement that multiplication and division operate only on ratio scales function as binding constraints within the national evaluative framework. The national-level focus establishes whether representational measurement principles are present as part of the epistemic infrastructure that governs quantitative claims regarding therapeutic value.

The logit profile demonstrates systematic exclusion of representational measurement axioms from the Czech Republic's national HTA knowledge base. Core statements asserting that measurement must precede arithmetic, that latent constructs require Rasch transformation to establish invariant interval scaling, and that multiplication requires ratio scale measurement collapse to floor or near-floor logit values. These results indicate non-possession of measurement principles as operational determinants within the national evaluative environment. Conversely, false statements asserting that QALYs constitute ratio measures, that composite preference-weighted indices can be validly multiplied and aggregated, and that simulation-based cost-effectiveness models generate empirically evaluable claims receive strong endorsement, reflected in high positive logit values. This pattern confirms that the Czech HTA framework operationalizes composite scoring constructs as if they were measurement-valid quantities. The national knowledge base therefore supports a structurally coherent administrative framework that produces quantitative outputs, but these outputs lack the measurement validity necessary to support lawful arithmetic, falsification, or empirical evaluation.

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 CZECH NATIONAL HTA KNOWLEDGE BASE

The Czech Republic’s national HTA knowledge base is structured around formal pharmacoeconomic evaluation requirements embedded within the reimbursement framework administered by the State Institute for Drug Control (Státní ústav pro kontrolu léčiv, SÚKL). Economic evaluation plays a central role in determining reimbursement eligibility, pricing, and access conditions for pharmaceuticals within the publicly funded health system. Manufacturers seeking reimbursement must submit economic evidence demonstrating comparative value, typically expressed through incremental cost-effectiveness analyses that employ cost-per-QALY metrics. These submissions integrate clinical trial evidence, epidemiological projections, and modeled estimates of long-term health outcomes, including survival and health-related quality of life.

At the national level, the Czech HTA framework reflects methodological convergence with broader European and international practice. Preference-based utility instruments, particularly the EQ-5D, are routinely used to derive utility weights representing health-related quality of life. These utility weights are combined with time to generate QALY estimates, which serve as the principal outcome measure in economic evaluation. Economic models, often structured as Markov simulations or decision analytic frameworks, project incremental costs and QALYs over extended time horizons. These modeled outputs form the quantitative basis for reimbursement decisions, pricing negotiations, and assessment of therapeutic value.

The national HTA knowledge base therefore rests on a methodological architecture that integrates clinical evidence, preference-based scoring systems, and simulation modeling. This structure provides administrative consistency, procedural transparency, and comparability across therapeutic interventions. Cost-effectiveness ratios derived from modeled QALYs function as integrative decision metrics, allowing comparative assessment across diverse disease areas and therapeutic classes. This framework aligns with international reference case standards and reflects diffusion of HTA methodology from organizations such as NICE and European collaborative networks.

However, the quantitative constructs embedded within this framework originate from composite utility scores derived from multidimensional health state classification systems. These scores represent aggregated ordinal preferences rather than invariant measures of a unidimensional attribute. Their numerical properties are determined by scoring algorithms and valuation conventions rather than by demonstration of measurement invariance or dimensional homogeneity. Despite this, arithmetic operations—including multiplication, division, and aggregation—are routinely performed on these scores to generate summary outcome measures and cost-effectiveness ratios.

Simulation modeling further extends the role of these constructs by projecting outcomes beyond observed clinical data. Models generate estimates of lifetime QALYs and associated costs,

producing numerical outputs that appear to quantify therapeutic impact. These outputs function as operational decision variables within the reimbursement framework. Yet the framework does not require demonstration that the underlying constructs satisfy the axioms of representational measurement, nor does it require transformation of subjective responses through Rasch measurement to establish invariant unit structure.

The Czech Republic's national HTA knowledge base therefore embodies a structured administrative framework that integrates quantitative constructs into decision making. It demonstrates procedural coherence and methodological consistency in applying established HTA conventions. However, it does not enforce the representational measurement conditions necessary to ensure that arithmetic operations performed within the framework correspond to empirical measurement. Composite utility indices function as operational decision variables despite lacking invariant unit properties, and simulation outputs provide numerical projections without empirical measurement foundation. The national HTA knowledge base thus reflects convergence with the global HTA methodological paradigm, in which quantitative sophistication coexists with the absence of measurement-valid constructs as the basis for evaluating therapy impact.

.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: CZECH NATIONAL HTA KNOWLEDGE BASE

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.

CZECH REPUBLIC: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT IN THE NATIONAL HTA KNOWLEDGE BASE

The Czech Republic's national HTA knowledge base, understood as the country's reimbursement assessment architecture, its pharmacoeconomic submission expectations, the methodological idiom that governs cost-utility claims, and the published justificatory literature around those practices exhibits the now-familiar signature of the global HTA memeplex: the axioms of representational measurement are not treated as preconditions for arithmetic, but as optional background that can be ignored while quantitative procedures proceed unimpeded (Table 1). The logit profile does not suggest occasional error at the margins; it depicts an evaluative culture in which foundational measurement propositions are structurally excluded as binding constraints while their negations are operationally affirmed. This inversion is not "academic." It is the condition that determines whether numbers used in pricing and access decisions correspond to measurable attributes or merely to administratively sanctioned scores.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS CZECH NATIONAL KNOWLEDGE BASE

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	-1.47
MEASURES MUST BE UNIDIMENSIONAL	1	0.10	-2.20
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+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	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.55	+0.52
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	0.06	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.50	+0.40
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

Two facts about the Czech setting matter immediately. First, Czech HTA in pharmaceuticals has been embedded in the reimbursement process since 2008, with SÚKL positioned centrally in the evaluation of medicines within formal administrative proceedings. Second, the economic evaluation idiom is cost-effectiveness and cost-utility, in which incremental ratios are routinely expressed in cost per QALY and compared to threshold or willingness-to-pay language attributed to methodological recommendations. Once that architecture is acknowledged, the logit pattern becomes predictable, because the framework cannot operate unless it assumes what measurement theory forbids.

Start with the "true" axioms: interval measures lack a true zero; unidimensionality is required; multiplication requires ratio measurement; measurement must precede arithmetic; and the axioms of representational measurement must be satisfied if arithmetic is to have empirical meaning. In the Czech national profile these propositions register at −1.40, −2.20, −2.20, −2.20, and −2.20 respectively. The interpretation is not subtle. These are floor-adjacent values indicating effective non-possession: within the Czech HTA knowledge base, these propositions do not function as rules that constrain what analysts are permitted to compute. They may be "known" in the trivial sense that an individual could acknowledge them if asked, but they are not binding. They do not operate. They do not discipline the analytical workflow. They have no enforcement mechanism, no institutional pathway, and no methodological status in the production of Czech HTA claims.

The reason is obvious: the Czech system, like NICE-derived systems globally, needs a single integrative number to close the evaluative case. It needs a scalar that can be multiplied, aggregated, ratio-compared, and thresholds to yield a decision endpoint. Cost-per-QALY is the canonical closure device. If the Czech environment were to accept, operationally, that multiplication requires

ratio measurement, and that composite ordinal preference scores cannot be "upgraded" by assertion to interval or ratio scales, then the reference case collapses. The logit evidence is therefore not merely descriptive; it is diagnostic of necessity. A knowledge base cannot operationalize cost-per-QALY closure while simultaneously honoring the axioms that would invalidate it.

This is why the "false" propositions, those that institutionalize arithmetic on non-measures, cluster strongly on the positive side. The QALY is treated as a ratio measure (+2.50). QALYs can be aggregated (+2.50). EQ-5D-type preference algorithms are treated as if they create interval measures (+2.20). Time trade-off preferences are treated as unidimensional (+1.75). Ratio measures can have negative values is endorsed as a denial (+2.20). Likert summation is treated as ratio by the denial of that claim (+2.50), and, critically, reference case simulations are treated as generating falsifiable claims by the denial of their falsifiability (+2.20). These positive logits do not mean the Czech HTA community has proved any of these things. They mean the Czech HTA knowledge base behaves as if they were true. They are treated as legitimate objects of calculation. They are embedded in the norms of publication, submission, and decision.

A QALY combines time and a utility weight. The Czech profile correctly recognizes time as ratio (+2.50). But this only highlights the asymmetry that defines HTA: lawful measurement is respected for manifest quantities when unavoidable, while being ignored for the latent constructs that the framework must manufacture to create an integrative decision number. The utility weight is not a manifest measure; it is a preference score derived from valuations over multiattribute health state descriptions. Nothing in the Czech operational architecture demonstrates that these preference scores form a unidimensional attribute with invariant unit structure. Yet they are treated as if they could be multiplied by time to create a quantity. Even if one were to grant interval properties to the preference scores, something that itself would require a lawful transformation framework for latent traits, the product still fails dimensional homogeneity. A quantity is not made coherent by multiplying unlike entities. "Utility-years" are not a homogeneous empirical attribute. They are an administrative hybrid: the number exists because the model needs it, not because the world contains it.

The Czech system illustrates the fundamental inversion found across countries: arithmetic is not the consequence of measurement; measurement is assumed as a post hoc justification for arithmetic. The logit profile makes this inversion explicit through the repeated suppression of "measurement precedes arithmetic" to −2.20. In a measurement-based science, the path is: define the attribute, establish its measurement structure, confirm admissible transformations, and only then apply arithmetic operations. In Czech HTA, as in the broader memeplex, the path is reversed: define the decision need (closure), select a computational framework to deliver closure, and then treat its outputs as if they were measures. The result is a stable administrative technology that produces numbers with policy authority but without empirical meaning. A memeplex of meaningless measurement.

Now consider the Rasch cluster. Four statements sit at the absolute floor: only linear ratio and Rasch logit ratio scales constitute valid measurement structures (−2.50); transforming ordinal responses to interval measurement is only possible with Rasch rules (−2.50); Rasch logit ratio scaling is the only basis for assessing therapy impact for latent traits (−2.50); and Rasch rules are

identical to the axioms of representational measurement (−2.50). This is decisive because it exposes the epistemic engine of the memeplex. If latent traits are to be quantified at all, the issue is not whether one prefers Rasch "psychometrics" versus "classical test theory." The issue is whether one can lawfully transform ordinal observations into a structure with invariant units. Rasch is not an option; it is the only coherent transformation model that attempts to satisfy those requirements for latent constructs. The Czech profile says that this knowledge does not exist as an operational constraint. It is outside the knowledge base as a binding methodology. The inevitable consequence is that latent constructs are never measured in Czech HTA. They are scored. They are indexed. They are normalized. They are modeled. But they are not measured.

This matters because Czech HTA decisions are not rhetorical exercises; they shape real access, real pricing, real restrictions, and real treatment availability. The Czech reimbursement environment is a formal administrative setting with measurable consequences for who receives therapies and under what conditions. When the quantitative constructs used to justify those decisions are not measurement-valid, the system cannot claim scientific accountability. It can claim procedural compliance. It can claim harmonization. It can claim transparency. But it cannot claim that its quantitative outputs represent therapy impact in a way that physicians and patients can rely on as empirically grounded.

Duty of care enters here in a precise way. A duty-of-care framework presupposes that claims about therapy impact are exposed to being wrong, because only then can decision makers learn, correct, and improve. The Czech profile's "non-falsifiable claims should be rejected" sits at +0.52—noticeably weaker than the strong endorsement of the false constructs. This asymmetry is revealing. The knowledge base can accommodate the rhetoric of falsifiability at the level of general principle, but it does not build falsifiability into the constructs that determine decisions. That is exactly what a memeplex does: it can repeat the language of science while operationally insulating itself from the risk of refutation.

Reference case simulation is the mechanism of insulation. Models can always be recalculated; they are rarely exposed to the binary discipline of being wrong. If one parameter changes, the output changes; if a new dataset appears, the model is updated; if a stakeholder objects, a scenario analysis is run. But none of this is falsification. It is iterative accommodation within a closed framework. The Czech profile's strong positive endorsement of "reference case simulations generate falsifiable claims" as a false statement (p=0.90; +2.20) captures the same pattern you have observed elsewhere: simulation outputs are treated as the apex of quantitative sophistication even though they are functions of assumptions, not observations. In the Czech environment, where cost/QALY idioms and WTP language appear in analyses and policy practice, the simulation becomes the bridge that allows a short-term clinical evidence base to be converted into long-term claims that cannot be directly observed. The bridge is computational, not empirical. It is therefore not testable in the sense required for the evolution of objective knowledge.

Once that is understood, the Czech profile can be read as a map of closure. The positive cluster is the closure cluster: QALYs as ratio; QALYs aggregable; preference algorithms producing interval; TTO unidimensional; composite subjective summations treated as arithmetic objects; simulations treated as legitimate decision machinery. The negative cluster is the openness cluster: measurement precedes arithmetic; unidimensionality; ratio requirements; representational axioms;

Rasch transformation; Rasch as the latent-trait measurement foundation. The openness cluster is excluded because openness destroys closure. A system that must produce a final cost-effectiveness claim will treat measurement axioms as obstacles, not as foundations.

This is why interpretive language such as "effective non-possession" is not merely rhetorical. At −2.50, the Rasch and "two measures" statements do not mean that Czech analysts have never heard the word Rasch, or that a Czech academic could not cite a psychometrics text. It means that within the boundary conditions of the national Czech HTA knowledge base guidance, submission norms, reimbursement practice, and the methodological literature that legitimizes it do not function as operative rules. They do not constrain acceptable claims. They are absent as determinants. That is the proper meaning of non-possession: not ignorance as a personal defect, but exclusion as a system property.

The Czech case also illustrates why "harmonization" in Europe can be an analytical dead end. If the Czech system aligns methodologically with the broader European cost-utility idiom, this is not an achievement of scientific convergence; it is a diffusion of the same measurement inversion. Harmonization makes it easier for manufacturers to prepare dossiers, easier for agencies to process them, and easier for journals to publish comparable analyses. But if what is harmonized is arithmetic on non-measures, then harmonization is simply the standardization of error. The Czech Republic's participation in this idiom should therefore be interpreted as administrative convergence, not measurement progress.

A defensible alternative exists and is simpler. Replace composite utility constructs with a portfolio of single-attribute claims. For manifest attributes, that means linear ratio measures expressed in units that are dimensionally homogeneous and interpretable (resource counts, event rates, days in hospital, time to event, avoidable admissions, and other unidimensional quantities). For latent constructs, it means Rasch logit ratio scaling with explicit protocols, invariant unit structure, and endpoints expressed as possession or change in possession on a defensible scale. Once claims are structured as single-attribute measures, falsification becomes possible: the claim specifies what should be observed in a defined target population within a defined timeframe, and the health system can replicate or refute it. That is how objective knowledge evolves. A cost/QALY framework cannot do this, because it cannot specify what it means to be wrong: it can always be re-parameterized, reweighted, or re-modeled.

The Czech national logit profile therefore belongs in the same category as those for other European Union countries. It demonstrates that Czech HTA, as presently constituted, is not a measurement-based enterprise. It is a numerical storytelling administrative system. It produces numbers with authority. It is capable of procedural sophistication. It can invoke "verifiable criteria" within administrative proceedings. But the verification is not measurement verification; it is procedural verification which is compliant with the required form of the dossier. The outcome is closure rather than learning. The cost-effectiveness claim is produced, the decision is made, and the next submission begins.

If the Czech system were to re-enter normal science, it would have to reverse the inversion: measurement first, arithmetic second; falsification as the criterion of progress; replication as the discipline of credibility; and duty of care as the ethical frame that requires claims to be empirically

evaluable rather than computationally persuasive. Until that reversal occurs, the Czech Republic's HTA knowledge base will remain a local instantiation of the global memeplex: a closed framework that treats composite scoring systems as if they were measures, treats simulation coherence as if it were empirical correspondence, and treats administratively convenient closure as if it were scientific discovery.

And that is the most damaging implication of the table: it is not that Czech HTA has made a few measurement errors. It is that the Czech national knowledge base, like the broader European and global corpus, is structured so that measurement-valid proposals—linear ratio measures for manifest claims and Rasch logit ratio measures for latent traits—are not evaluated, not debated, and not rejected. They fail to register as legitimate. They sit outside what the system recognizes as "quantification." That is exactly how memeplexes survive: by making the foundations of their own refutation invisible.

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