MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

POLAND: AOTMiT ENDORSEMENT OF MEASUREMENT FAILURE IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 86 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The Agency for Health Technology Assessment and Tariff System (AOTMiT) is Poland's national body responsible for supporting reimbursement and pricing decisions within the publicly funded healthcare system. It operates under the authority of the Ministry of Health and provides formal recommendations concerning the inclusion of pharmaceuticals, medical devices, and health services in reimbursement schemes.

AOTMiT's responsibilities include evaluation of clinical effectiveness, comparative analysis of therapeutic alternatives, economic evaluation, and assessment of budget impact. The agency reviews manufacturer submissions and prepares appraisal reports that inform ministerial decisions on reimbursement level, price negotiations, and coverage conditions. In addition to pharmaceutical assessments, AOTMiT plays a role in tariff setting for healthcare services through its involvement in cost analysis and pricing methodologies used within the national health insurance system.

The agency also develops and maintains methodological guidance specifying requirements for HTA submissions, including preferred analytic approaches, modeling expectations, and outcome measures. Through these guidelines and its assessment practice, AOTMiT helps define the technical standards governing health technology evaluation in Poland. AOTMiT therefore functions both as an evaluator of individual technologies and as a central institutional authority shaping how evidence and numerical claims are generated, interpreted, and accepted within the Polish health system.

The objective of this study is to examine whether the national health technology assessment knowledge base in Poland, as operationalized through the activities and guidance of the Agency for Health Technology Assessment and Tariff System (AOTMiT), possesses the conceptual conditions required for quantitative measurement. Rather than evaluating individual submissions, analyst competence, or methodological execution, the analysis interrogates the epistemic environment that authorizes numerical claims within Polish HTA practice. Using a twenty-four item canonical diagnostic derived from representational measurement theory and Rasch principles, the study seeks to determine whether foundational axioms governing scale type, unidimensionality, invariance, falsifiability, and the logical precedence of measurement over arithmetic function as operative admissibility constraints within the national HTA system.

The findings demonstrate a coherent and highly structured endorsement profile that mirrors those observed previously in other national HTA systems. Foundational propositions required for measurement consistently exhibit strong negative reinforcement, indicating that measurement axioms do not operate as governing rules within the Polish HTA knowledge base. Conversely, propositions known to be false under representational measurement theory—but necessary for valuation-based economic modeling—receive positive reinforcement. This pattern does not reflect inconsistency, transition, or methodological confusion. It reveals structural non-possession of measurement theory, combined with systematic normalization of numerical practice in the absence of representational validity. Poland’s HTA system therefore aligns not only procedurally but epistemically with international HTA frameworks, reinforcing the conclusion that contemporary HTA has converged globally on valuation-based numeracy rather than on measurement science.

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 AOTMiT KNOWLEDGE BASE

The knowledge base of the AOTMiT in Poland cannot be understood solely through its statutory mandate, organizational structure, or formal procedural guidance. Like all mature health technology assessment systems, its epistemic authority does not arise from a single document or institutional act. It arises from a distributed corpus of texts, practices, assumptions, and routines that collectively authorize numerical claims within decision making. The AOTMiT knowledge base is therefore not reducible to agency reports alone. It is an epistemic system.

This system includes methodological guidelines issued by AOTMiT, appraisal reports prepared by assessment teams, economic evaluation templates required for manufacturer submissions, and the technical conventions governing cost-effectiveness analysis, budget impact modeling, and comparative effectiveness assessment. Together, these documents define what constitutes acceptable evidence, which numerical outputs are admissible, and how competing claims are to be evaluated. Importantly, this authority is exercised not through explicit theoretical argument, but through procedural specification. What must be submitted, modeled, or reported becomes what is treated as legitimate.

At the core of this knowledge base lies the routine use of cost-utility analysis, typically structured around quality-adjusted life years derived from preference-based health-related quality of life instruments such as the EQ-5D. These numerical outputs are treated as quantitative measures suitable for arithmetic operations, including aggregation across individuals, multiplication by time, and comparison through incremental cost-effectiveness ratios. Their admissibility is established administratively rather than epistemically. They are accepted because they conform to established methodological templates, not because their measurement properties have been demonstrated.

The AOTMiT knowledge base also incorporates a substantial modeling infrastructure. Reference-case economic models, sensitivity analyses, scenario analyses, and probabilistic simulations are treated as central evidentiary tools. These models are evaluated primarily on internal coherence, transparency, and compliance with guidance rather than on the representational validity of their underlying variables. Numerical uncertainty is explored extensively, while the ontological status of the numbers themselves remains unexamined. In this way, uncertainty analysis substitutes for measurement validation.

Academic health economics literature plays an important reinforcing role. Polish HTA practice draws heavily on international methodological norms disseminated through journals, training materials, and European HTA collaboration frameworks. These sources transmit established conventions regarding utilities, QALYs, and modeling without interrogating their measurement foundations. As a result, methodological legitimacy is inherited rather than demonstrated. International alignment functions as epistemic justification.

Education and professional training further stabilize this structure. Analysts are trained to implement accepted techniques rather than to question the conditions under which numerical representation is valid. Students learn how to calculate utilities, populate models, and interpret thresholds, but are not taught to examine scale types, unidimensionality, invariance, or the logical relationship between measurement and arithmetic. By the time analysts enter professional roles, numerical legitimacy has already been internalized as common sense.

Crucially, the AOTMiT knowledge base does not contain explicit engagement with representational measurement theory. Concepts such as admissible arithmetic operations, true zero, invariant units, or latent trait possession do not function as governing constraints within assessment practice. Their absence is not the result of rejection. They simply do not appear as part of the evaluative vocabulary. Where axioms are not recognized, they cannot discipline inference.

The result is an epistemic environment in which numerical practice is normalized through repetition rather than justified through theory. Once utilities, QALYs, and modeled outcomes are routinely used, their quantitative status becomes self-confirming. Numbers are treated as measures because they are used as measures. Over time, this circular reinforcement produces epistemic closure: the foundational question of whether these numbers measure anything at all no longer arises as a legitimate scientific problem.

The AOTMiT knowledge base therefore operates as a coherent and stable system of numerical authorization. It does not misunderstand measurement theory; it does not possess it. Quantitative claims are permitted not because representational conditions have been satisfied, but because institutional consensus has rendered such questions invisible. This structural non-possession defines the epistemic limits within which Polish health technology assessment currently operates.

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: AOTMiT POLAND

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.

POLAND: REJECTING REPRESENTATIONAL MEASUREMENT FOR NUMERICAL STORYTELLING

The application of the 24-item canonical diagnostic to the AOTMiT knowledge base (Table 1) reveals a pattern that is not accidental, transitional, or ambiguous. The endorsement structure exhibits a coherent and internally consistent epistemic architecture in which numerical practice is authorized independently of representational measurement constraints. The pattern is stable across foundational axioms, scale-type requirements, latent-trait conditions, and falsifiability standards. What emerges is not a series of methodological oversights, but a systematic absence of measurement as a governing principle; an endorsement of numerical storytelling.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS AOTMiT POLAND

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	0.15	-1.75
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	0.85	+1.75
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.80	+1.40
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	0.90	+2.0
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

QALYS CAN BE AGGREGATED	0	0.90	+2.20
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.30	-0.95
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.40	-0.45
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	0.80	+1.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.05	-2.50
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The first group of statements interrogates the most elementary requirements of quantitative representation. These include the absence of a true zero on interval scales, the necessity of unidimensionality, the restriction of multiplication to ratio measures, and the principle that measurement must logically precede arithmetic. Across these propositions, endorsement probabilities are uniformly low, producing strongly negative logits. The results indicate that these axioms do not function as admissibility conditions within the AOTMiT knowledge base. They are not reinforced, taught, or operationalized as constraints on numerical reasoning.

This absence is not partial. The logits do not cluster around neutrality. They collapse decisively toward the lower bound of the scale. Such collapse is diagnostic. It indicates non-possession rather than disagreement. The knowledge base does not contest these axioms; it simply does not contain them as operative rules. Arithmetic is therefore permitted to proceed without prior establishment of measurement.

This epistemic ordering becomes explicit in the treatment of time trade-off preferences and QALYs. The statements asserting that time trade-off preferences are unidimensional and that the QALY constitutes a ratio measure receive strong positive endorsement. These are not marginal assumptions. They are foundational commitments required for cost-utility analysis to function. Their strong positive logits indicate that they are normalized as valid quantitative premises within AOTMiT-guided assessment practice.

Yet under representational measurement theory, both propositions are false. Preferences are ordinal judgments of desirability, not manifestations of a latent quantitative attribute. The QALY, constructed through multiplication of time by ordinal utilities lacking a true zero, cannot satisfy ratio properties. The positive endorsement of these false propositions therefore signals not error, but necessity. In the absence of measurement, valuation assumptions must be accepted in order for arithmetic to proceed at all. The inversion becomes clearer when examining statements concerning negative values. The proposition that ratio measures can have negative values receives one of the strongest positive endorsements in the table. This is epistemically decisive. Negative values are incompatible with ratio scales by definition, as ratio scales require an absolute zero representing absence of the attribute. The acceptance of negative utilities therefore represents an explicit abandonment of measurement constraints in favor of valuation logic.

This is not a technical oversight. It is a categorical choice. The knowledge base privileges preference representation over attribute representation. Once that choice is made, negative values are not problematic; they are required. Without them, certain trade-off constructions collapse. The positive logit therefore marks structural dependence rather than conceptual confusion.

The same logic governs the treatment of aggregation. Statements asserting that QALYs can be aggregated and that summated subjective responses create ratio measures receive strong positive endorsement. Aggregation is indispensable to population-level modeling, budget impact analysis, and comparative assessment. Without aggregation, HTA cannot function administratively. As a result, propositions that enable aggregation are normalized regardless of their incompatibility with measurement theory.

By contrast, the statement asserting that claims for cost-effectiveness fail the axioms of representational measurement collapses to the floor of the scale. This is particularly revealing. It demonstrates that the knowledge base does not recognize cost-effectiveness modeling as epistemically contingent upon measurement validity. Cost-effectiveness is treated as a legitimate numerical construct by procedural conformity rather than by representational grounding.

The diagnostic thus exposes the epistemic hierarchy governing AOTMiT assessment practice. Arithmetic necessity overrides measurement admissibility. What can be modeled is treated as what can be measured. This reversal is fundamental. Under representational measurement theory, arithmetic is permitted only after measurement has been established. Within the AOTMiT knowledge base, measurement is presumed because arithmetic is required.

This inversion becomes even more pronounced in the treatment of latent constructs. Statements asserting that Rasch transformation is necessary for interval measurement of latent traits, that the logit ratio scale is the only lawful basis for latent trait comparison, and that the outcome of interest for latent traits is possession rather than score magnitude all collapse to the lowest logit values. These propositions are not weakly endorsed. They are absent.

The implication is stark. Although AOTMiT assessments routinely invoke constructs such as health-related quality of life, functioning, or utility, the knowledge base does not contain the conceptual apparatus required to measure latent traits. Rasch measurement does not function as an admissibility rule. Invariance is not required. Possession is not defined. Scores are treated as quantities by convention alone.

This is not ignorance in the ordinary sense. It is structural non-possession. Measurement theory does not operate anywhere within the epistemic workflow. It is not invoked in guidelines, not referenced in evaluation criteria, not required in submissions, and not used to adjudicate competing claims. Where axioms are absent, they cannot be selectively violated. They simply do not exist.

The pattern extends to falsifiability. The statement that non-falsifiable claims should be rejected receives only weak endorsement, while the claim that reference case simulations generate falsifiable claims receives strong positive endorsement. This pairing is revealing. It indicates that model outputs are treated as empirically testable by virtue of internal coherence rather than external measurement linkage. Simulation becomes evidence by construction.

From a philosophy-of-science perspective, this represents a decisive departure from normal science. Falsifiability requires that claims be capable of empirical refutation through observation. Simulation outputs cannot meet this requirement when their parameters lack measurement validity. Yet within the AOTMiT knowledge base, falsifiability is redefined procedurally. A model is treated as testable if alternative scenarios can be generated, not if empirical measurement can adjudicate its claims.

The endorsement of the logit definition itself illustrates the surface-level nature of quantitative literacy within the system. The proposition defining the logit as the natural logarithm of the odds ratio receives moderate endorsement. This suggests technical familiarity without epistemic integration. The mathematics is recognized, but its measurement implications are not. Knowing what a logit is does not translate into requiring logit scales for latent trait measurement.

Taken together, the 24-item pattern reveals a closed epistemic system. The knowledge base is not internally contradictory. On the contrary, it is highly coherent. All propositions that would constrain numerical practice are absent. All propositions required to enable numerical practice are endorsed. The resulting structure is stable precisely because it excludes the conditions that could destabilize it. This coherence explains why incremental reform cannot succeed. Adjusting thresholds, refining models, or updating value sets does not alter the underlying epistemic ordering. Measurement cannot be introduced downstream. It must exist at the point where numbers first claim to represent attributes. Once that boundary is crossed without measurement, all subsequent quantitative reasoning inherits the same defect.

The canonical diagnostic therefore does not identify isolated weaknesses in AOTMiT practice. It identifies the epistemic architecture within which that practice operates. The agency’s methods are consistent with European HTA norms not because they are independently justified, but because those norms share the same structural commitments. Arithmetic is authorized by consensus, not by representation. This finding also clarifies why AOTMiT’s position is not exceptional. The endorsement profile aligns closely with those observed in Finland, the United Kingdom, Canada, and Australia. The convergence is not empirical coincidence. It reflects epistemic inheritance of false measurement. These systems did not independently fail measurement; they inherited a valuation-based ontology in which numbers precede attributes.

The significance of this result lies in its diagnostic clarity. It removes discretion from interpretation. The issue is not whether AOTMiT applies HTA methods competently. It does. The issue is whether those methods rest on measurable quantities. The diagnostic shows they do not. Numerical precision within such a system is therefore illusory. Confidence intervals surround non-quantities. Incremental ratios compare magnitudes that have not been established. Apparent rigor masks representational absence. This is not pseudoscience in the sense of error or fraud. It is something more entrenched: a numerical belief system that has substituted valuation for measurement so completely that the distinction is no longer visible.

Recognizing this does not require abandoning HTA. It requires confronting its boundary conditions. Either measurement is required for quantitative claims, or it is not. If it is required, then the current AOTMiT framework cannot support those claims. If it is not required, then HTA must relinquish the language of magnitude, precision, and empirical inference. It must recognize itself as pseudo-science; outside the standards of normal science for falsification and the evolution of objective knowledge. The 24-item canonical diagnostic makes that choice unavoidable. It shows that AOTMiT does not lack better models or more refined instruments. It lacks measurement as an admissibility condition. Until that condition is restored, numerical storytelling will remain structurally invariant, regardless of how sophisticated the analysis appears.

POLAND: IS AOTMiT IN A POSITION TO PROPOSE PRODUCT ASSESSMENT GUIDELINES THAT MEET REPRESENTATIONAL MEASUREMENT STANDARDS?

The question of whether AOTMiT is in a position to propose product assessment guidelines that meet the axioms of representational measurement is not primarily a question of legal authority, but of institutional function. Although AOTMiT participates directly in reimbursement recommendations, its central role within the Polish health system is epistemic rather than decisional. It evaluates evidence, structures assessment standards, and defines the numerical forms through which value claims are expressed. It does not itself determine prices, negotiate access agreements, or control final ministerial decisions. This distinction is decisive.

Because AOTMiT functions as the methodological gateway through which product claims must pass, it occupies a structurally privileged position within the HTA architecture. While reimbursement decisions rest with the Ministry of Health, the epistemic legitimacy of those decisions depends entirely on the analytical framework supplied by AOTMiT. In this sense, AOTMiT does not merely apply methodology; it defines what counts as admissible evidence in the first place.

This positioning matters because representational measurement standards cannot be introduced through incremental modification of existing cost-utility frameworks. They require explicit recognition that numerical claims must satisfy admissibility conditions prior to arithmetic. That recognition entails abandoning the assumption that preference-weighted multiattribute indices generate quantitative magnitudes. For decision-making authorities, such acknowledgment is institutionally destabilizing. It would call into question decades of precedent, thresholds, and comparative claims. For AOTMiT, however, the risk is qualitatively different.

AOTMiT is not required to defend historical reimbursement outcomes. Its mandate is prospective: to assess submissions under defined methodological rules. This grants it a degree of epistemic freedom unavailable to ministries, payers, or pricing committees. It can revise standards governing future submissions without invalidating past decisions. This temporal asymmetry creates a narrow but genuine opportunity for reform.

From the standpoint of representational measurement theory, the core requirement is straightforward: numerical claims may be admitted only when the numbers preserve empirically testable relational structure. This implies that multiattribute preference-based utilities cannot be treated as measures of magnitude, regardless of their international prevalence. Arithmetic cannot be justified by convention. Measurement must precede modeling, not follow it.

AOTMiT is institutionally capable of stating this principle without prescribing immediate substitutes. Measurement theory does not require agencies to specify instruments. It requires them to define admissibility. AOTMiT could therefore distinguish clearly between three categories of evidence: descriptive classification, preference research, and measurement-based outcome claims. Each may retain value, but only the latter would be eligible for arithmetic operations such as aggregation, multiplication by time, or incremental comparison.

Such a guideline would not prohibit the use of EQ-5D, utilities, or QALYs. It would reclassify them. Preference-based indices could continue to inform deliberation, contextual interpretation, and narrative comparison, while being explicitly excluded from claims of quantitative magnitude. Measurement claims, by contrast, would be restricted to outcomes satisfying scale-type requirements: linear ratio measures for manifest attributes and Rasch logit ratio measures for latent constructs.

Crucially, this approach aligns with AOTMiT’s existing role. It would not require the agency to endorse a specific outcome framework, disease model, or instrument family. Representational measurement theory is content-neutral. It governs structure, not substance. AOTMiT could therefore frame revised guidelines as a pre-analytic filter governing numerical legitimacy rather than as a competing methodological doctrine.

The advantage of this strategy is institutional coherence. Rather than increasing technical complexity, it restores conceptual clarity. Analysts would remain free to innovate, model, and explore uncertainty, but arithmetic would be permitted only where quantities exist. This shifts the center of HTA from modeling ingenuity to evidentiary legitimacy.

Importantly, such guidelines could be explicitly prospective. Their purpose would not be to invalidate existing submissions or overturn historical recommendations, but to establish measurement-compliant standards for future product assessment. This avoids institutional paralysis while acknowledging that the current numerical architecture lacks foundational justification.

The alternative is continued epistemic drift. Ongoing harmonization with European HTA conventions that conflate valuation with measurement simply reproduces structural non-measurement under the banner of alignment. In that environment, convergence becomes replication of error rather than correction of it.

AOTMiT is therefore not merely capable of proposing representationally valid product assessment guidelines; it occupies one of the few positions within European HTA where such reform is structurally possible. Its authority lies precisely in defining evidence rather than enforcing outcomes. Because it governs admissibility rather than access, it can confront foundational questions that decision-making bodies cannot.

Whether AOTMiT chooses to exercise this capacity is a policy decision. Institutionally, however, it stands at the only point in the Polish HTA system where measurement can be reintroduced as a governing principle without destabilizing the entire evaluative architecture. If representational measurement is to become an admissibility condition for quantitative claims, AOTMiT is the logical and perhaps the only locus from which that transition could occur.

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.

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