MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

FINLAND: FINCCHTA 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 81 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The Finnish Coordinating Center for Health Technology Assessment (FINCCHTA) occupies a distinctive position within the Finnish health system. Unlike jurisdictions with centralized reimbursement authorities, Finland's HTA structure is decentralized, distributed across government agencies, hospital districts, and academic research units. FINCCHTA was established to provide coordination, methodological alignment, and knowledge integration across this fragmented environment. Its role is not primarily regulatory, but epistemic.

FINCCHTA functions as a national hub through which health technology assessment methods, standards, and evaluative conventions are consolidated and disseminated. It supports HTA activity undertaken by universities, hospital districts, and public agencies by promoting common methodological frameworks, facilitating information exchange, and aligning Finnish practice with European HTA collaboration initiatives. In doing so, it exerts influence not by issuing binding decisions, but by shaping what counts as acceptable evidence and legitimate analytical practice.

This coordinating role gives FINCCHTA particular epistemic significance. While it does not determine reimbursement outcomes directly, it helps define the numerical language through which technologies are assessed. Instruments, outcome measures, modeling approaches, and interpretive norms that circulate through FINCCHTA-supported assessments acquire institutional legitimacy. Over time, this legitimacy becomes self-reinforcing. Methods endorsed through coordination become standard practice, and standard practice becomes assumed correctness.

FINCCHTA's influence is further amplified by Finland's integration into European HTA networks. Participation in joint assessments and methodological harmonization initiatives has encouraged alignment with dominant European conventions, particularly those relying on preference-based quality-of-life instruments and cost-utility frameworks. In this context, FINCCHTA operates as a conduit through which international HTA norms are translated into national practice.

The importance of FINCCHTA for the present analysis therefore lies not in any single guideline or report, but in its systemic function. As a coordinating body, it shapes the epistemic environment in which numerical claims are produced, interpreted, and accepted. The question addressed in this working paper is not whether FINCCHTA applies accepted methods correctly, but whether the methodological environment it sustains recognizes the conditions under which numerical outcomes may legitimately be treated as measures.

Understanding FINCCHTA in this way is essential. The findings that follow do not describe failure by individual analysts or institutions. They describe the epistemic consequences of coordination itself when harmonization proceeds without foundational measurement constraints.

The objective of this study is to examine whether FINCCHTA operates within an epistemic framework consistent with the axioms of representational measurement. Rather than evaluating specific reimbursement decisions, analytical techniques, or modeling practices, the analysis interrogates the conceptual foundations that authorize numerical claims within the Finnish health technology assessment environment. The central question is not whether FINCCHTA applies accepted methods competently, but whether the knowledge base that governs its evaluative practice recognizes the conditions under which numerical outcomes may legitimately represent empirical quantities.

To address this question, the study applies the 24-item canonical measurement diagnostic developed in the Logit Working Papers series. This diagnostic draws explicitly on representational measurement theory and Rasch principles to distinguish between propositions that must be true for quantitative measurement to occur and propositions that are known to be false yet are required for preference-based utility systems to function. Endorsement probabilities and normalized logits are used to assess whether these propositions operate as governing constraints within the FINCCHTA epistemic environment. The purpose is diagnostic rather than evaluative: to determine whether measurement theory is possessed as an admissibility framework or absent as a conceptual authority.

The results demonstrate that FINCCHTA exhibits a stable and internally coherent pattern of measurement non-possession. Foundational axioms of representational measurement—including unidimensionality, invariant units, scale-type restrictions, and the logical precedence of measurement over arithmetic receive uniformly low endorsement probabilities, generating strongly negative logit values. These principles do not function as operative constraints within the Finnish HTA knowledge base. Measurement theory does not appear as a governing rule set capable of disciplining numerical interpretation.

Conversely, propositions that are incompatible with measurement theory but necessary for preference-based utility modeling, such as the treatment of multiattribute utilities as quantitative measures, the acceptance of negative values on purported ratio scales, aggregation of utilities, and arithmetic manipulation of preference scores—receive strong positive endorsement. This polarity is not contradictory but structural. The knowledge base consistently rejects the axioms that would prohibit arithmetic while affirming the assumptions that permit it. The resulting epistemic configuration mirrors patterns observed in other national HTA systems and indicates not methodological confusion but a coherent commitment to valuation-based numerical practice. FINCCHTA therefore operates within a system that produces numbers without measurement and treats arithmetic as valid in the absence of representational justification.

The starting point is simple and inescapable: *measurement precedes arithmetic*. This principle is not a methodological preference but a logical necessity. One cannot multiply what one has not measured, cannot sum what has no dimensional homogeneity, cannot compare ratios when no ratio scale exists. When HTA multiplies time by utilities to generate QALYs, it is performing arithmetic with numbers that cannot support the operation. When HTA divides cost by QALYs, it is constructing a ratio from quantities that have no ratio properties. When HTA aggregates QALYs across individuals or conditions, it is combining values that do not share a common scale. These practices are not merely suboptimal; they are mathematically impossible.

The modern articulation of this principle can be traced to Stevens' seminal 1946 paper, which introduced the typology of nominal, ordinal, interval, and ratio scales [1]. Stevens made explicit what physicists, engineers, and psychologists already understood: different kinds of numbers permit different kinds of arithmetic. Ordinal scales allow ranking but not addition; interval scales permit addition and subtraction but not multiplication; ratio scales alone support multiplication, division, and the construction of meaningful ratios. Utilities derived from multiattribute preference exercises, such as EQ-5D or HUI, are ordinal preference scores; they do not satisfy the axioms of interval measurement, much less ratio measurement. Yet HTA has, for forty years, treated these utilities as if they were ratio quantities, multiplying them by time to create QALYs and inserting them into models without the slightest recognition that scale properties matter. Stevens' paper should have blocked the development of QALYs and cost-utility analysis entirely. Instead, it was ignored.

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of Measurement* (1971) [2] . Representational Measurement Theory (RMT) formalized the axioms under which empirical attributes can be mapped to numbers in a way that preserves structure. Measurement, in this framework, is not an act of assigning numbers for convenience, it is the discovery of a lawful relationship between empirical relations and numerical relations. The axioms of additive conjoint measurement, homogeneity, order, and invariance specify exactly when interval scales exist. RMT demonstrated once and for all that measurement is not optional and not a matter of taste: either the axioms hold and measurement is possible, or the axioms fail and measurement is impossible. Every major construct in HTA, utilities, QALYs, DALYs, ICERs, incremental ratios, preference weights, health-state indices, fails these axioms. They lack unidimensionality; they violate independence; they depend on aggregation of heterogeneous attributes; they collapse under the requirements of additive conjoint measurement. Yet HTA proceeded, decade after decade, without any engagement with these axioms, as if the field had collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

Whereas representational measurement theory articulates the axioms for interval measurement, Georg Rasch's 1960 model provides the only scientific method for transforming ordered categorical responses into interval measures for latent traits [3]. Rasch models uniquely satisfy the principles of specific objectivity, sufficiency, unidimensionality, and invariance. For any construct such as pain, fatigue, depression, mobility, or need, Rasch analysis is the only legitimate means of producing an interval scale from ordinal item responses. Rasch measurement is not an alternative to RMT; it is its operational instantiation. The equivalence of Rasch's axioms and the axioms of representational measurement was demonstrated by Wright, Andrich and others as early as the 1970s. In the latent-trait domain, the very domain where HTA claims to operate; Rasch is the only game in town [4].

Yet Rasch is effectively absent from all HTA guidelines, including NICE, PBAC, CADTH, ICER, SMC, and PHARMAC. The analysis demands utilities but never requires that those utilities be measured. They rely on multiattribute ordinal classifications but never understand that those constructs be calibrated on interval or ratio scales. They mandate cost-utility analysis but never justify the arithmetic. They demand modelled QALYs but never interrogate their dimensional

properties. These guidelines do not misunderstand Rasch; they do not know it exists. The axioms that define measurement and the model that makes latent trait measurement possible are invisible to the authors of global HTA rules. The field has evolved without the science that measurement demands.

How did HTA miss the bus so thoroughly? The answer lies in its historical origins. In the late 1970s and early 1980s, HTA emerged not from measurement science but from welfare economics, decision theory, and administrative pressure to control drug budgets. Its core concern was *valuing health states*, not *measuring health*. This move, quiet, subtle, but devastating, shifted the field away from the scientific question "What is the empirical structure of the construct we intend to measure?" and toward the administrative question "How do we elicit a preference weight that we can multiply by time?" The preference-elicitation projects of that era (SG, TTO, VAS) were rationalized as measurement techniques, but they never satisfied measurement axioms. Ordinal preferences were dressed up as quasi-cardinal indices; valuation tasks were misinterpreted as psychometrics; analyst convenience replaced measurement theory. The HTA community built an entire belief system around the illusion that valuing health is equivalent to measuring health. It is not.

The endurance of this belief system, forty years strong and globally uniform, is not evidence of validity but evidence of institutionalized error. HTA has operated under conditions of what can only be described as *structural epistemic closure*: a system that has never questioned its constructs because it never learned the language required to ask the questions. Representational measurement theory is not taught in graduate HTA programs; Rasch modelling is not part of guideline development; dimensional analysis is not part of methodological review. The field has been insulated from correction because its conceptual foundations were never laid. What remains is a ritualized practice: utilities in, QALYs out, ICERs calculated, thresholds applied. The arithmetic continues because everyone assumes someone else validated the numbers.

This Logit Working Paper series exposes, through probabilistic and logit-based interrogations of AI large language national knowledge bases, the scale of this failure. The results display a global pattern: true statements reflecting the axioms of measurement receive weak endorsement; false statements reflecting the HTA belief system receive moderate or strong reinforcement. This is not disagreement. It is non-possession. It shows that HTA, worldwide, has developed as a quantitative discipline without quantitative foundations; a confused exercise in numerical storytelling.

The conclusion is unavoidable: HTA does not need incremental reform; it needs a scientific revolution. Measurement must precede arithmetic. Representational axioms must precede valuation rituals. Rasch measurement must replace ordinal summation and utility algorithms. Value claims must be falsifiable, protocol-driven, and measurable; rather than simulated, aggregated, and numerically embellished.

The global system of non-measurement is now visible. The task ahead is to replace it with science.

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

This analysis is generated through the structured interrogation of a large language model (LLM) applied to a defined documentary corpus and is intended solely to characterize patterns within an aggregated knowledge environment. It does identify, assess, or attribute beliefs, intentions, competencies, or actions to any named individual, faculty member, student, administrator, institution, or organization. The results do not constitute factual findings about specific persons or programs, nor should they be interpreted as claims regarding professional conduct, educational quality, or compliance with regulatory or accreditation standards. All probabilities and logit values reflect model-based inferences about the presence or absence of concepts within a bounded textual ecosystem, not judgments about real-world actors. The analysis is exploratory, interpretive, and methodological in nature, offered for scholarly discussion of epistemic structures rather than evaluative or legal purposes. Any resemblance to particular institutions or practices is contextual and non-attributive, and no adverse implication should be inferred.

1. INTERROGATING THE LARGE LANGUAGE MODEL

A large language model **(**LLM**)** is an artificial intelligence system designed to understand, generate, and manipulate human language by learning patterns from vast amounts of text data. Built on deep neural network architectures, most commonly transformers, LLMs analyze relationships between words, sentences, and concepts to produce contextually relevant responses. During training, the model processes billions of examples, enabling it to learn grammar, facts, reasoning patterns, and even subtle linguistic nuances. Once trained, an LLM can perform a wide range of tasks: answering questions, summarizing documents, generating creative writing, translating languages, assisting with coding, and more. Although LLMs do not possess consciousness or true understanding, they simulate comprehension by predicting the most likely continuation of text based on learned patterns. Their capabilities make them powerful tools for communication, research, automation, and decision support, but they also require careful oversight to ensure accuracy, fairness, privacy, and responsible use

In this Logit Working Paper, "interrogation" refers not to discovering what an LLM *believes,* it has no beliefs, but to probing the content of the *corpus-defined knowledge space* we choose to analyze. This knowledge base is enhanced if it is backed by accumulated memory from the user. In this case the interrogation relies also on 12 months of HTA memory from continued application of the system to evaluate HTA experience. The corpus is defined before interrogation: it may consist of a journal (e.g., *Value in Health*), a national HTA body, a specific methodological framework, or a collection of policy documents. Once the boundaries of that corpus are established, the LLM is used to estimate the conceptual footprint within it. This approach allows us to determine which principles are articulated, neglected, misunderstood, or systematically reinforced.

In this HTA assessment, the objective is precise: to determine the extent to which a given HTA knowledge base or corpus, global, national, institutional, or journal-specific, recognizes and reinforces the foundational principles of representational measurement theory (RMT). The core principle under investigation is that measurement precedes arithmetic; no construct may be treated as a number or subjected to mathematical operations unless the axioms of measurement are satisfied. These axioms include unidimensionality, scale-type distinctions, invariance, additivity, and the requirement that ordinal responses cannot lawfully be transformed into interval or ratio quantities except under Rasch measurement rules.

The HTA knowledge space is defined pragmatically and operationally. For each jurisdiction, organization, or journal, the corpus consists of:

- published HTA guidelines
- agency decision frameworks
- cost-effectiveness reference cases
- academic journals and textbooks associated with HTA
- modelling templates, technical reports, and task-force recommendations
- teaching materials, methodological articles, and institutional white papers

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The "knowledge base" is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

THE NATIONAL KNOWLEDGE BASE FOR FINCCHTA

The FINCCHTA knowledge base cannot be understood as a discrete set of methodological documents or technical manuals. It is best conceptualized as an epistemic corpus: a distributed body of texts, practices, conventions, and institutional routines through which numerical authority is established and maintained. This corpus includes methodological guidance documents, commissioned HTA reports, academic collaborations, training materials, conference proceedings, and the routine use of internationally standardized outcome measures within Finnish evaluations.

Central to this knowledge base is the integration of Finland into the broader European HTA environment. FINCCHTA has historically aligned its methodological practices with European coordination initiatives, including EUnetHTA and related joint assessment frameworks. This alignment privileges comparability, transferability, and procedural harmonization. Numerical outcomes are expected to travel across institutional contexts, and instruments such as preference-based quality-of-life measures are adopted primarily for their compatibility with international practice rather than for their measurement properties.

Within this environment, numerical outputs are treated as self-evidently quantitative. Utilities, QALYs, and modeled outcomes are routinely interpreted as magnitudes capable of comparison, aggregation, and trade-off. The epistemic legitimacy of these numbers does not derive from explicit demonstration of scale properties. Instead, it arises through repetition and institutional acceptance. Once numerical forms become embedded in assessment templates and reporting standards, their quantitative status is assumed rather than examined.

Importantly, the FINCCHTA knowledge base does not articulate measurement theory as a governing framework. Concepts such as unidimensionality, invariant units, true zero, or admissible arithmetic operations do not appear as threshold criteria for numerical claims. Their absence does not reflect explicit rejection. Rather, they are not part of the evaluative grammar through which HTA reasoning is conducted. As a result, numerical practice proceeds without encountering representational constraints.

Education and professional training reinforce this structure. Analysts are taught how to apply established instruments, populate models, and interpret outputs within accepted HTA conventions. They are not trained to interrogate whether the numbers they manipulate constitute measures in the representational sense. By the time analysts enter practice, numerical legitimacy has already been internalized as a methodological given.

The authority of the knowledge base is therefore circular but stable. Developers point to international norms. Agencies point to precedent. Analysts point to guidance. Each component

defers foundational justification elsewhere, resulting in epistemic closure. Measurement theory remains external to the system and therefore incapable of disciplining it.

In this sense, the FINCCHTA knowledge base does not merely permit false measurement; it normalizes it. Numerical claims are sustained not through representational validity but through institutional coherence. The system functions smoothly, consistently, and transparently—yet without measurement as an admissibility condition. This structural absence, rather than any technical deficiency, defines the epistemic character of Finnish health technology assessment.

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: FINCCHTA

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

It is essential to understand how to interpret these results.

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

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

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

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

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS FINCCHTA

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	0.10	-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	0.85	+1.75
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.85	+1.75
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	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50
QALYS CAN BE AGGREGATED	0	0.90	+2.20
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1		
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

FINCCHTA: REPRESENTATIONAL MEASUREMENT WAS NEVER AN ISSUE

The canonical interrogation of FINCCHTA reveals a knowledge base that is not transitional, conflicted, or epistemically uncertain. Rather, it displays a stable and internally coherent pattern of non-measurement that mirrors the endorsement profiles observed in other national HTA agencies examined in the Logit Working Papers series. The Finnish case is particularly instructive, not because it deviates from international norms, but because it demonstrates how deeply measurement failure has become institutionalized even within research cultures that otherwise emphasize methodological rigor, transparency, and analytic sophistication.

At the most foundational level, FINCCHTA does not recognize the axioms of representational measurement as admissibility conditions for quantitative claims. This absence is immediately evident in the endorsement probabilities associated with the most elementary measurement propositions. The requirement that measures be unidimensional receives an endorsement probability of only 0.10, corresponding to a normalized logit of −2.20. This indicates near-total absence of reinforcement. Unidimensionality does not function as a governing rule within the FINCCHTA epistemic environment. It is neither articulated nor enforced, despite being a necessary condition for any quantity claim under representational measurement theory.

This absence is decisive. Without unidimensionality, there is no empirical attribute whose magnitude can be meaningfully represented. Numerical aggregation across heterogeneous domains cannot produce quantity; it produces indexation. Yet within FINCCHTA-supported assessments, multiattribute instruments are routinely treated as if they measured a single underlying construct. The diagnostic reveals that this treatment is not supported by any recognition of measurement axioms. It persists by convention alone.

The same pattern appears with respect to scale type. The proposition that multiplication requires a ratio measure also receives an endorsement probability of 0.10 (−2.20). This indicates that scale-

type distinctions do not operate as constraints on numerical practice. Whether an outcome may be multiplied by time, aggregated across persons, or entered into cost-effectiveness ratios is treated as a modeling decision rather than a measurement question. Arithmetic is permitted not because it is justified, but because it is required by evaluative frameworks.

Closely related is the proposition that measurement must precede arithmetic. Here again, the endorsement probability collapses to 0.10. This finding is central. In the FINCCHTA knowledge base, arithmetic does not follow measurement; it substitutes for it. Numbers become meaningful through use rather than through representational justification. Once embedded within accepted HTA routines, numerical manipulation is assumed to confer legitimacy retrospectively.

The requirement that arithmetic must satisfy the axioms of representational measurement fares no better. With an endorsement probability of 0.10, this proposition does not function as an admissibility condition. FINCCHTA documentation, methodological guidance, and accepted analytic practice do not require demonstration of scale validity prior to arithmetic operations. This omission is not accidental. It reflects an epistemic environment in which measurement theory is simply absent as a category of reasoning.

The most decisive results emerge with respect to Rasch measurement. All Rasch-related propositions collapse uniformly to the floor of the scale, with endorsement probabilities of 0.05 and normalized logits of −2.50. The knowledge base does not recognize Rasch transformation as necessary for latent trait measurement. It does not recognize the logit ratio scale as the only defensible basis for quantifying latent attributes. It does not recognize equivalence between Rasch axioms and representational measurement theory. These principles do not operate at all.

This is not a matter of disagreement or methodological preference. Rasch measurement is not rejected; it is invisible. It does not appear as a requirement, a reference point, or a constraint. Consequently, latent constructs such as health-related quality of life are treated as if they were directly measurable through preference scoring, despite lacking invariant units, unidimensional structure, or separability of person and item parameters.

The proposition that the outcome of interest for latent traits is possession of that trait likewise collapses to −2.50. This finding reveals a fundamental conceptual shift within the FINCCHTA epistemic system. Health outcomes are not conceptualized as attributes possessed by individuals. Instead, individuals are located within valued health states. The numerical output reflects how those states are judged, not how much of an attribute an individual possesses.

This distinction is foundational. Measurement concerns properties of entities. Valuation concerns preferences of observers. The FINCCHTA knowledge base does not preserve this distinction. It allows valuations of hypothetical states to stand in for magnitudes of individual health. Once this substitution occurs, the entire edifice of quantitative inference becomes detached from empirical structure.

While the failure to reinforce true measurement axioms is severe, it is only half of the epistemic pattern. The other half emerges in the strong positive endorsement of propositions that are known

to be false under representational measurement theory but are necessary for utility-based HTA to function.

The proposition that ratio measures can have negative values receives an endorsement probability of 0.90, corresponding to a normalized logit of +2.20. This indicates strong and stable reinforcement. Negative utilities are not treated as anomalous or problematic; they are normalized. Yet under representational measurement theory, negative values are incompatible with ratio scales, which require a true zero representing absence of the attribute. The coexistence of negative values and ratio arithmetic constitutes a direct contradiction. FINCCHTA resolves this contradiction not by addressing it, but by ignoring it.

Similarly, the claim that the QALY is a ratio measure receives strong positive endorsement (+2.20). This acceptance is not accompanied by any demonstration that the QALY satisfies ratio axioms. Instead, ratio status is assumed because multiplication by time is required. Arithmetic necessity substitutes for measurement justification.

The proposition that QALYs can be aggregated also clusters at +2.20. Aggregation across individuals is treated as routine. Yet aggregation presupposes dimensional homogeneity, invariant units, and additive structure. None of these conditions are established. The diagnostic reveals that aggregation is authorized by convention alone.

Preference-based scoring is likewise normalized. The propositions that summation of subjective responses creates ratio measures and that Likert-type scores can be summed both receive strong positive logits. These endorsements do not indicate belief in their truth; they indicate their indispensability. Without accepting these propositions, the numerical infrastructure of HTA would collapse. Their reinforcement is therefore constitutive, not accidental.

The proposition that EQ-5D preference algorithms create interval measures also receives strong positive endorsement, even though no theoretical basis exists for this claim. Preferences express order and desirability, not magnitude. Yet the FINCCHTA knowledge base treats valuation outputs as if they possessed interval properties sufficient for subtraction and comparison. The distinction between preference and quantity is erased.

The result is a complete epistemic inversion. Propositions that should function as constraints are absent. Propositions that should be prohibited are normalized. Measurement axioms disappear, while their negation becomes routine practice. The probability–logit structure captures this inversion with striking clarity.

Notably, this profile is not internally contradictory. It is coherent. FINCCHTA does not oscillate between incompatible positions. It consistently endorses a framework in which numerical legitimacy derives from institutional acceptance rather than from representational validity. The knowledge base is stable precisely because it does not contain competing measurement principles that could generate tension.

This coherence explains the resilience of the Finnish HTA system to critique. Because measurement axioms do not function as admissibility conditions, challenges framed in

measurement terms fail to gain traction. There is no internal language with which to evaluate them. Arguments about unidimensionality, invariance, or scale type appear irrelevant because the system does not recognize them as governing rules.

The endorsement probability for the rejection of non-falsifiable claims provides further insight. With a probability of 0.30 (−0.95), FINCCHTA exhibits weak and inconsistent reinforcement of falsification as a requirement. This reflects the broader HTA reliance on simulation modeling. Reference-case models are accepted despite being structurally non-falsifiable. Their outputs are treated as evidence even though no empirical test could refute them. This is not perceived as a problem because the purpose of the model is alignment, not discovery.

Correspondingly, the proposition that reference case simulations generate falsifiable claims receives strong positive endorsement. This reflects not misunderstanding but institutional necessity. Simulation outputs must be treated as falsifiable for the system to function, regardless of whether they actually are.

Together, these results show that FINCCHTA operates within a fully developed numerical belief system. It is not partially scientific. It is not methodologically immature. It is epistemically complete; but complete in non-measurement. Importantly, this analysis does not imply incompetence or bad faith. FINCCHTA analysts are highly trained. Their methods are sophisticated. Their documentation is transparent. The failure identified here is not technical. It is foundational. The system does not recognize measurement theory as binding.

This distinction between ignorance and non-possession is crucial. Measurement theory has been available for more than a century. The axioms articulated by Stevens, Suppes, Krantz, Luce, and Tukey were well established before modern HTA emerged. Rasch measurement has been available since 1960. The absence observed here cannot be attributed to novelty or obscurity. It reflects deliberate epistemic inheritance. HTA did not grow out of measurement science. It grew out of decision science. Its objective was not to discover quantities but to support choices. Valuation frameworks were therefore adopted early because they facilitated comparison, not because they measured attributes. Once institutionalized, these frameworks shaped training, publication norms, and methodological guidance. Measurement theory was never incorporated, and over time it became invisible.

The FINCCHTA diagnostic confirms that this inheritance remains intact. Despite national differences in governance and culture, the Finnish HTA knowledge base reproduces the same epistemic structure observed elsewhere. Measurement does not precede arithmetic. Valuation substitutes for measurement. Simulation substitutes for falsification. The implication is unavoidable. FINCCHTA does not merely permit false measurement. It institutionalizes it. By accepting numerical claims unsupported by representational measurement axioms, it confers legitimacy on arithmetic that lacks empirical foundation. This legitimacy then propagates downstream through pricing negotiations, reimbursement decisions, and policy discourse.

This does not mean that FINCCHTA decisions are arbitrary or unethical. It means that the numerical claims used to support them do not constitute measurements. They are structured expressions of preference and assumption, not representations of magnitude. No amount of

refinement can correct this condition. Adjusting tariffs, expanding descriptive systems, or improving modeling techniques cannot introduce unidimensionality, invariant units, or true zero where none exist. Measurement cannot be added downstream. It must exist upstream.

The canonical interrogation therefore yields a clear conclusion. FINCCHTA functions within an epistemic system that does not recognize the axioms of representational measurement as admissibility conditions for quantitative claims. Its endorsement profile is not anomalous. It is exemplary. Finland has not failed to measure. It has never attempted to measure.

IS FINCCHTA IN A POSTION TO PROPOSE A PRODUCT ASSESMENT GUIDELIE THAT MEETS REPRESENTATIONAL MEASREMENT STANDARDS

The question of whether FINCCHTA is in a position to propose product assessment guidelines that meet the axioms of representational measurement is not a question of authority, but of institutional function. FINCCHTA does not operate as a reimbursement decision maker, nor does it control pricing or access directly. Its role is epistemic rather than coercive. Precisely for that reason, it is uniquely positioned to initiate methodological reform that would be difficult for decision-making agencies to undertake.

Because FINCCHTA functions as a coordinating center rather than a gatekeeping authority, it is not institutionally bound to defend legacy numerical practices. Agencies responsible for reimbursement are structurally constrained: their decisions depend on continuity, precedent, and procedural stability. Acknowledging foundational measurement failure would destabilize prior determinations and expose decisions to retrospective challenge. FINCCHTA faces no such constraint. Its mandate is not to justify past decisions, but to guide future assessment practice.

This distinction matters. Representational measurement standards cannot be introduced through incremental adjustment to 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 produce quantities. For reimbursement agencies, this admission is institutionally threatening. For a coordinating methodological body, it is not.

FINCCHTA therefore occupies a rare epistemic position. It can propose assessment guidelines that separate measurement from valuation without immediately disrupting access decisions. It can distinguish descriptive classification, preference research, and measurement-based outcome claims as categorically different forms of evidence. In doing so, it can restore conceptual clarity without requiring the health system to renounce decision making altogether.

The essential requirement would be to redefine what counts as an admissible quantitative claim in product assessment. Under representational measurement theory, numerical outputs are admissible only when they preserve empirically testable relational structure. This implies that claims based on multiattribute utility indices cannot be treated as measures of magnitude, regardless of how widely they are used internationally. FINCCHTA is institutionally capable of stating this without dictating what must replace them immediately.

Such guidelines would not prohibit descriptive instruments or preference studies. They would reclassify them. Preference-based outputs could be retained as contextual or deliberative information, but explicitly excluded from arithmetic operations such as aggregation, comparison of magnitude, or time-based multiplication. Measurement claims, by contrast, would be restricted to outcomes that satisfy scale-type requirements: linear ratio measures for manifest attributes and Rasch logit ratio measures for latent attributes.

Importantly, this does not require FINCCHTA to endorse a specific instrument or outcome set. Representational measurement theory does not prescribe content. It prescribes admissibility. FINCCHTA could therefore frame its guidelines not as a competing methodology, but as a pre-analytic filter governing which numerical claims may legitimately support inference. This preserves methodological pluralism while restoring logical discipline.

The advantage of such an approach is that it aligns with FINCCHTA's coordinating mandate. Rather than issuing prescriptive templates, it would define boundaries. Analysts could continue to innovate within those boundaries, but arithmetic would be permitted only where measurement exists. This shifts the center of gravity of assessment from modeling sophistication to evidentiary legitimacy.

Critically, FINCCHTA could position these guidelines as prospective rather than corrective. The objective would not be to invalidate existing submissions or past analyses, but to establish standards for future product assessment. This avoids institutional paralysis while acknowledging that the current numerical architecture lacks measurement foundations.

The alternative is epistemic drift. Continued alignment with European HTA conventions that conflate valuation with measurement entrenches structural non-measurement as a permanent feature of Finnish assessment practice. In that environment, harmonization becomes a mechanism for reproducing error rather than correcting it.

FINCCHTA is therefore not merely capable of proposing representationally valid product assessment guidelines; it may be one of the few institutions in Europe structurally able to do so. Its authority lies precisely in not deciding. Because it coordinates knowledge rather than enforcing outcomes, it can confront foundational questions that decision-making bodies cannot.

Whether FINCCHTA chooses to exercise this capacity is a separate matter. But institutionally, it is positioned at the only point in the Finnish HTA system where measurement can be reintroduced without destabilizing the entire evaluative architecture. If representational measurement is to be restored as an admissibility condition for quantitative claims, it is difficult to identify a more appropriate locus for that transition.

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