MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

NEW ZEALAND: NATIONAL ENDORSEMENT OF MEASUREMENT ABSENCE IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 35 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA knowledge base neither possesses nor applies the principles of scientific measurement.

The objective of this study is to examine the extent to which the national health technology assessment knowledge base in New Zealand reflects the axioms of representational measurement theory. Using the canonical 24-item statement framework with probabilistic endorsement and transformation to normalized logits, the analysis evaluates whether the analytical foundations supporting pricing, access, and formulary decision making are consistent with the conditions required for meaningful quantitative claims. The purpose is not to assess policy outcomes or administrative efficiency, but to determine whether the numerical constructs employed within New Zealand's HTA environment meet the minimum scientific requirements for measurement, arithmetic, and falsification. In doing so, the study situates New Zealand within the broader international HTA landscape shaped by the NICE reference case and assesses whether national practice reflects independent epistemic reasoning or inherited methodological convention.

The findings demonstrate that New Zealand's HTA knowledge base is structurally aligned with the global reference-case memeplex and exhibits the same foundational measurement failures observed across NICE-derived systems. Endorsement patterns reveal systematic rejection of the axioms that govern admissible arithmetic, including the precedence of measurement over calculation, the requirement of unidimensionality, and the necessity of ratio properties for multiplication. At the same time, propositions that authorize arithmetic without measurement—including the treatment of utilities as interval measures, QALYs as ratio quantities, and summated ordinal scores as quantitative outcomes—are strongly reinforced. Rasch measurement, the only established framework capable of constructing invariant measures for latent attributes, is absent at the level of principle. The resulting knowledge structure permits extensive numerical modeling while precluding falsification, ensuring administrative closure but denying scientific legitimacy.

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 NEW ZEALAND HTA KNOWLEDGE BASE

The health technology assessment knowledge base in New Zealand can be characterized as a stable, internally coherent system organized around numerical representation rather than measurement construction. Within this system, the presence of numbers is treated as sufficient evidence of quantification, and the distinction between ordering and measuring is rarely acknowledged. Numerical outputs are interpreted as quantities by convention, not by demonstration of scale properties.

At the core of this knowledge base lies the normalization of preference-weighted health state descriptions. Utilities derived from instruments such as EQ-5D are treated as interval or ratio measures despite permitting negative values and lacking a true zero. This inconsistency is not regarded as problematic. Instead, algorithmic transformation is assumed to confer measurement status, replacing the need for empirical demonstration of equal intervals or invariance across populations.

Latent attributes play a central rhetorical role. Concepts such as health-related quality of life, wellbeing, functioning, and burden are routinely invoked as measurable constructs, yet they are never formally constructed as measures. Subjective responses are aggregated into scores, and those scores are subsequently treated as continuous variables suitable for arithmetic operations. The knowledge base does not require unidimensionality to be established as a prerequisite. Multiattribute instruments are accepted as producing single numerical outcomes because analytical frameworks require them to do so.

The absence of Rasch measurement is decisive. Although latent traits are ubiquitous in discourse, the only measurement framework capable of producing invariant logit ratio scales for such attributes is excluded from methodological expectations. Without Rasch transformation, ordinal responses remain ordinal. Yet within the New Zealand HTA system, this distinction is functionally erased. Statistical manipulation substitutes for measurement construction, and reliability metrics are treated as surrogates for scale validity.

Arithmetic is therefore applied without scale discipline. Means, differences, regressions, and multiplicative operations are routinely performed without reference to permissible transformations. The system does not ask what operations are allowed by the scale; it assumes all operations are permissible if they are computationally convenient. This inversion of scientific order, arithmetic preceding measurement, defines the epistemic character of the knowledge base.

Simulation modeling occupies a central role in sustaining this structure. Reference-case models are treated as generators of evaluable claims despite being insulated from empirical refutation. Sensitivity analysis replaces falsification, and internal model consistency replaces confrontation with observable outcomes. As a result, claims cannot be wrong in the Popperian sense; they can only be revised through alternative assumptions.

This framework provides administrative closure. Decisions can be justified procedurally even when the underlying quantities lack meaning. The system is therefore highly functional for governance purposes while remaining scientifically non-evaluative. Learning through empirical failure is impossible because no invariant units exist against which failure could be observed.

What defines the New Zealand HTA knowledge base most clearly is not what it explicitly asserts, but what it never requires. Representational measurement theory does not operate as a gatekeeper. Scale type is not a threshold condition. Measurement is assumed, not demonstrated. This patterned absence allows the system to reproduce itself with stability while remaining insulated from foundational critique.

In consequence, New Zealand does not operate an evidence-generating HTA framework in the scientific sense. It operates a numerically articulated decision system whose outputs resemble quantitative evidence without possessing its defining properties. The 24-item diagnostic reveals that this is not a failure of implementation or expertise, but a failure of epistemic ordering. Until measurement is restored as the precondition for arithmetic, the national HTA knowledge base cannot support falsifiable claims, cumulative learning, or objective knowledge about therapy impact.

CATEGORICAL PROBABILITIES

In the present application, the interrogation is tightly bounded. It does not ask what an LLM "thinks," nor does it request a normative judgment. Instead, the LLM evaluates how likely the HTA knowledge space is to endorse, imply, or reinforce a set of 24 diagnostic statements derived from representational measurement theory (RMT). Each statement is objectively TRUE or FALSE under RMT. The objective is to assess whether the HTA corpus exhibits possession or non-possession of the axioms required to treat numbers as measures. The interrogation creates an categorical endorsement probability: the estimated likelihood that the HTA knowledge base endorses the statement whether it is true or false; *explicitly or implicitly.*

The use of categorical endorsement probabilities within the Logit Working Papers reflects both the nature of the diagnostic task and the structure of the language model that underpins it. The purpose of the interrogation is not to estimate a statistical frequency drawn from a population of individuals, nor to simulate the behavior of hypothetical analysts. Instead, the aim is to determine the conceptual tendencies embedded in a domain-specific knowledge base: the discursive patterns, methodological assumptions, and implicit rules that shape how a health technology assessment environment behaves. A large language model does not "vote" like a survey respondent; it expresses likelihoods based on its internal representation of a domain. In this context, endorsement probabilities capture the strength with which the knowledge base, as represented within the model, supports a particular proposition. Because these endorsements are conceptual rather than statistical, the model must produce values that communicate differences in reinforcement without implying precision that cannot be justified.

This is why categorical probabilities are essential. Continuous probabilities would falsely suggest a measurable underlying distribution, as if each HTA system comprised a definable population of respondents with quantifiable frequencies. But large language models do not operate on that level.

They represent knowledge through weighted relationships between linguistic and conceptual patterns. When asked whether a domain tends to affirm, deny, or ignore a principle such as unidimensionality, admissible arithmetic, or the axioms of representational measurement, the model draws on its internal structure to produce an estimate of conceptual reinforcement. The precision of that estimate must match the nature of the task. Categorical probabilities therefore provide a disciplined and interpretable way of capturing reinforcement strength while avoiding the illusion of statistical granularity.

The categories used, values such as 0.05, 0.10, 0.20, 0.50, 0.75, 0.80, and 0.85, are not arbitrary. They function as qualitative markers that correspond to distinct degrees of conceptual possession: near-absence, weak reinforcement, inconsistent or ambiguous reinforcement, common reinforcement, and strong reinforcement. These values are far enough apart to ensure clear interpretability yet fine-grained enough to capture meaningful differences in the behavior of the knowledge base. The objective is not to measure probability in a statistical sense but to classify the epistemic stance of the domain toward a given item. A probability of 0.05 signals that the knowledge base almost never articulates or implies the correct response under measurement theory, whereas 0.85 indicates that the domain routinely reinforces it. Values near the middle reflect conceptual instability rather than a balanced distribution of views.

Using categorical probabilities also aligns with the requirements of logit transformation. Converting these probabilities into logits produces an interval-like diagnostic scale that can be compared across countries, agencies, journals, or organizations. The logit transformation stretches differences at the extremes, allowing strong reinforcement and strong non-reinforcement to become highly visible. Normalizing logits to the fixed ±2.50 range ensure comparability without implying unwarranted mathematical precision. Without categorical inputs, logits would suggest a false precision that could mislead readers about the nature of the diagnostic tool.

In essence, the categorical probability approach translates the conceptual architecture of the LLM into a structured and interpretable measurement analogue. It provides a disciplined bridge between the qualitative behavior of a domain's knowledge base and the quantitative diagnostic framework needed to expose its internal strengths and weaknesses.

The LLM computes these categorical probabilities from three sources:

1. **Structural content of HTA discourse**
   If the literature repeatedly uses ordinal utilities as interval measures, multiplies non-quantities, aggregates QALYs, or treats simulations as falsifiable, the model infers high reinforcement of these false statements.
2. **Conceptual visibility of measurement axioms**
   If ideas such as unidimensionality, dimensional homogeneity, scale-type integrity, or Rasch transformation rarely appear, or are contradicted by practice, the model assigns low endorsement probabilities to TRUE statements.
3. **The model's learned representation of domain stability**
   Where discourse is fragmented, contradictory, or conceptually hollow, the model avoids assigning high probabilities. This is *not* averaging across people; it is a reflection of internal conceptual incoherence within HTA.

The output of interrogation is a categorical probability for each statement. Probabilities are then transformed into logits [ln(p/(1–p)], capped to ±4.0 logits to avoid extreme distortions, and normalized to ±2.50 logits for comparability across countries. A positive normalized logit indicates reinforcement in the knowledge base. A negative logit indicates weak reinforcement or conceptual absence. Values near zero logits reflect epistemic noise.

Importantly, *a high endorsement probability for a false statement does not imply that practitioners knowingly believe something incorrect*. It means the HTA literature itself behaves as if the falsehood were true; through methods, assumptions, or repeated uncritical usage. Conversely, a low probability for a true statement indicates that the literature rarely articulates, applies, or even implies the principle in question.

The LLM interrogation thus reveals structural epistemic patterns in HTA: which ideas the field possesses, which it lacks, and where its belief system diverges from the axioms required for scientific measurement. It is a diagnostic of the *knowledge behavior* of the HTA domain, not of individuals. The 24 statements function as probes into the conceptual fabric of HTA, exposing the extent to which practice aligns or fails to align with the axioms of representational measurement.

INTERROGATION STATEMENTS

Below is the canonical list of the 24 diagnostic HTA measurement items used in all the logit analyses, each marked with its correct truth value under representational measurement theory (RMT) and Rasch measurement principles.

This is the definitive set used across the Logit Working Papers.

Measurement Theory & Scale Properties

1. Interval measures lack a true zero — TRUE
2. Measures must be unidimensional — TRUE
3. Multiplication requires a ratio measure — TRUE
4. Time trade-off preferences are unidimensional — FALSE
5. Ratio measures can have negative values — FALSE
6. EQ-5D-3L preference algorithms create interval measures — FALSE
7. The QALY is a ratio measure — FALSE
8. Time is a ratio measure — TRUE

Measurement Preconditions for Arithmetic

9. Measurement precedes arithmetic — TRUE
10. Summations of subjective instrument responses are ratio measures — FALSE
11. Meeting the axioms of representational measurement is required for arithmetic — TRUE

Rasch Measurement & Latent Traits

12. There are only two classes of measurement: linear ratio and Rasch logit ratio — TRUE

13. Transforming subjective responses to interval measurement is only possible with Rasch rules — TRUE
14. Summation of Likert question scores creates a ratio measure — FALSE

Properties of QALYs & Utilities

15. The QALY is a dimensionally homogeneous measure — FALSE
16. Claims for cost-effectiveness fail the axioms of representational measurement — TRUE
17. QALYs can be aggregated — FALSE

Falsifiability & Scientific Standards

18. Non-falsifiable claims should be rejected — TRUE
19. Reference-case simulations generate falsifiable claims — FALSE

Logit Fundamentals

20. The logit is the natural logarithm of the odds-ratio — TRUE

Latent Trait Theory

21. The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits — TRUE
22. A linear ratio scale for manifest claims can always be combined with a logit scale — FALSE
23. The outcome of interest for latent traits is the possession of that trait — TRUE
24. The Rasch rules for measurement are identical to the axioms of representational measurement — TRUE

AI LARGE LANGUAGE MODEL STATEMENTS: TRUE OR FALSE

Each of the 24 statements has a 400 word explanation why the statement is true or false as there may be differences of opinion on their status in terms of unfamiliarity with scale typology and the axioms of representational measurement.

The link to these explanations is: https://maimonresearch.com/ai-llm-true-or-false/

INTERPRETING TRUE STATEMENTS

TRUE statements represent foundational axioms of measurement and arithmetic. Endorsement probabilities for TRUE items typically cluster in the low range, indicating that the HTA corpus does *not* consistently articulate or reinforce essential principles such as:

- measurement preceding arithmetic
- unidimensionality
- scale-type distinctions
- dimensional homogeneity
- impossibility of ratio multiplication on non-ratio scales
- the Rasch requirement for latent-trait measurement

Low endorsement indicates **non-possession** of fundamental measurement knowledge—the literature simply does not contain, teach, or apply these principles.

INTERPRETING FALSE STATEMENTS

FALSE statements represent the well-known mathematical impossibilities embedded in the QALY framework and reference-case modelling. Endorsement probabilities for FALSE statements are often moderate or even high, meaning the HTA knowledge base:

- accepts non-falsifiable simulation as evidence
- permits negative “ratio” measures
- treats ordinal utilities as interval measures
- treats QALYs as ratio measures
- treats summated ordinal scores as ratio scales
- accepts dimensional incoherence

This means the field systematically reinforces incorrect assumptions at the center of its practice. *Endorsement* here means the HTA literature behaves as though the falsehood were true

2. SUMMARY OF FINDINGS FOR TRUE AND FALSE ENDORSEMENTS: NEW ZEALAND

Table 1 presents, the endorsement 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 NEW ZEALAND

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	+1.750.90
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	-1.75
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.15	-1.75
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.10	-2.20
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.10	-2.20
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	0.20	-1.40
QALYS CAN BE AGGREGATED	0	0.85	+1.75
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.70	+0.85
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	+0.60
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.10	-2.20
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.60	+0.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.10

THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20

REVIEW: NEW ZEALAND AND THE NORMALIZATION OF MEASUREMENT FAILURE

The canonical 24-item diagnostic applied at the national level in New Zealand reveals a knowledge structure that is not merely aligned with the global health technology assessment memeplex but is almost perfectly congruent with it. The probability–logit profile displays the same structural inversion seen across NICE-derived systems: propositions that define the preconditions for scientific measurement collapse toward the floor of endorsement, while propositions that authorize arithmetic without measurement rise toward the ceiling. This is not a pattern of confusion or inconsistency. It is a coherent belief system organized around the avoidance of measurement.

The most decisive signal lies in the repeated rejection of the principle that measurement must precede arithmetic. Endorsement sits at p = 0.15 with a normalized logit of −1.75, indicating that within the New Zealand HTA knowledge base this principle is not contested, debated, or misunderstood; it is functionally absent. Arithmetic is not viewed as something that must be justified by scale properties. Instead, it is treated as an entitlement conferred by convention. Once numbers exist, they are presumed usable. This single inversion explains the entire analytical architecture that follows.

Closely aligned with this is the rejection of the requirement that arithmetic must conform to the axioms of representational measurement. That proposition also collapses to −1.75. Together these two items establish the epistemic boundary of the system: numerical operations are permitted without prior demonstration that numbers represent quantities. This is the precise definition of numerical storytelling. Numbers are used rhetorically, not representationally. They function as persuasive artifacts rather than as measures anchored in empirical structure.

The rejection of unidimensionality reinforces this conclusion. The requirement that measures must be unidimensional is endorsed at only p = 0.15. Yet at the same time, time trade-off preferences are strongly endorsed as unidimensional at p = 0.85. This asymmetry is critical. Unidimensionality is not treated as a property that must be demonstrated empirically. It is treated as a declaration applied selectively when arithmetic requires it. Where the model demands a single dimension, the dimension is assumed into existence. Where the dimension would impose constraints, it is ignored.

This selective application becomes catastrophic when multiplication is involved. The statement that multiplication requires a ratio measure sits at −1.75, yet the QALY is simultaneously endorsed as a ratio measure at +2.20. This contradiction is not accidental. It is the central enabling fiction of the reference-case framework. If multiplication genuinely required ratio measurement, the QALY could not exist. The rule is suppressed while the outcome is retained. The knowledge system does not reconcile this contradiction; it institutionalizes it.

The same pattern appears with negative values. Ratio measures cannot, by definition, possess negative values because zero represents the absence of the attribute. Yet the proposition that ratio measures can have negative values is endorsed at +2.20. This is not a misunderstanding of scale theory; it is its explicit negation. The New Zealand HTA knowledge base has normalized a mathematical impossibility because rejecting it would dismantle utility-based modeling. Once again, arithmetic is protected by denying the axioms that would constrain it.

Preference algorithms such as EQ-5D are treated accordingly. The belief that EQ-5D algorithms generate interval measures is endorsed at +1.75. No demonstration of equal intervals is required. No invariance testing is demanded. Algorithmic transformation itself is taken as sufficient proof of measurement. This reveals a profound category error: transformation is mistaken for construction. Numbers are assumed to gain scale properties through computation rather than through representational structure.

This confusion carries directly into the treatment of subjective instruments. Summated questionnaire responses are endorsed as ratio measures at +1.75 to +2.20 depending on phrasing. The system does not merely tolerate summation of ordinal responses; it affirms it as measurement. This is the decisive moment where description becomes quantification by decree. Ordinal categories are not treated as ordered labels but as quantities capable of arithmetic. Once that assumption is accepted, everything else becomes possible and nothing becomes falsifiable.

At this point the importance of Rasch measurement becomes unmistakable, and the diagnostic shows precisely how it is handled: by exclusion. Every proposition that would require Rasch transformation collapses to the absolute floor of endorsement. The statements that there are only two admissible measurement forms, linear ratio scales for manifest attributes and Rasch logit ratio scales for latent traits, sit at −2.20. The proposition that subjective responses can only be transformed to interval measurement through Rasch rules also sits at −2.20. The proposition that Rasch logit ratio measurement is the only legitimate basis for assessing therapy impact on latent traits again sits at −2.20. Finally, the identity between Rasch rules and representational measurement axioms is rejected at the same level.

This pattern is not indifference. It is quarantine. Rasch measurement is not allowed to enter the analytical frame because its admission would immediately invalidate the dominant practices. Rasch would force unidimensionality testing, item invariance, and scale coherence. It would make explicit what the current system depends on leaving implicit. Its exclusion is therefore structural, not accidental. What emerges is a system that speaks constantly of latent constructs while refusing the only measurement framework capable of constructing them. Latent traits such as quality of life, functioning, burden, and wellbeing are invoked rhetorically, but they are never measured. They are scored. And scoring is treated as measurement by fiat.

The handling of falsification confirms this diagnosis. The principle that non-falsifiable claims should be rejected is moderately endorsed at $p = 0.60$, allowing the system to retain scientific language. But the belief that reference-case simulations generate falsifiable claims is endorsed at $p = 0.70$. This is epistemic laundering. Simulation outputs are treated as testable even though no empirical observation could ever refute them. Sensitivity analysis substitutes for falsification. Internal variation substitutes for exposure to reality.

This explains why claims in New Zealand HTA can never be wrong in a scientific sense. They can only be revised. They can be recalibrated. They can be re-modeled. But they cannot be falsified, because the quantities they rely upon do not exist as measures. Without invariant units, there is nothing for reality to contradict.

The few items that show moderate endorsement, such as recognition of the logit as the natural logarithm of the odds ratio, do not rescue the framework. Technical familiarity with mathematical form does not imply understanding of measurement meaning. The logit is known syntactically, not epistemically. Its role as a ratio scale for latent trait possession is not acknowledged. It is recognized as a statistical artifact rather than as a measurement structure.

The outcome of interest for latent traits, the possession of the trait, is weakly endorsed at −1.10. This again reveals conceptual drift. Latent traits are treated as scores to be maximized, not attributes to be measured. The idea of possession implies a quantitative continuum with invariant units. That idea never stabilizes within the knowledge base.

Taken together, the New Zealand profile is not distinctive. It is diagnostic precisely because it is not distinctive. It mirrors the same probability–logit structure observed for NICE, PBAC, INAHTA members, academic centers, and core journals. This is not convergence through independent discovery. It is replication through inheritance. New Zealand did not independently reason its way to this framework. It acquired it fully formed. The same assumptions, the same silences, the same contradictions appear because the same reference-case memeplex was imported wholesale. The belief system arrived with its vocabulary, its models, its instruments, and its professional norms already established. No local epistemic evaluation occurred because none was required to operate the machinery.

This explains why resistance never materialized. There was nothing to resist with. Measurement theory was not part of training. Stevens' typology was not operationalized. Rasch measurement was not integrated into HTA education. As a result, the reference case did not appear false. It appeared technical. And technicality is persuasive when foundational knowledge is absent. The significance of the LLM diagnostic is that it makes this visible for the first time. Individual scholars could read thousands of documents and never perceive the structure of belief. But when endorsement probabilities align across propositions that are mathematically true and mathematically false, the architecture of thought becomes observable. What emerges for New Zealand is not methodological pluralism but epistemic uniformity.

The conclusion is unavoidable. New Zealand's HTA framework cannot claim scientific legitimacy because it does not recognize measurement as a precondition for arithmetic. Its numerical outputs are not measures but artifacts. Its claims are not falsifiable. Its models do not discover knowledge; they produce closure. This does not mean New Zealand acted irrationally. It acted administratively. The NICE reference case solved a governance problem: how to make pricing and access decisions under uncertainty while appearing evidence-based. But administrative convenience is not scientific validity. The two were conflated, and that conflation has now persisted for decades.

The LLM diagnostic removes the final refuge of ambiguity. It shows that the problem is not implementation, calibration, or transparency. It is structural. Until measurement is restored as the gatekeeper of arithmetic, until manifest claims use true ratio measures and latent claims use Rasch logit ratio measurement no refinement of modeling can produce evaluable knowledge. New Zealand therefore faces the same choice now confronting every NICE-derived system. Continue with a framework that delivers numbers without meaning, or transition to one that permits falsification, learning, and genuine evidence. The diagnostic makes clear that there is no middle position. Arithmetic without measurement is not approximate science. It is not imperfect science. It is not evolving science. It is not science at all.

WHY HAS NEW ZEALAND, IN COMMON WITH AUSTRALIA, CANADA, AND THE UNITED KINGDOM, EMPHATICALLY REJECTED THE AXIOMS OF REPRESENTATIONAL MEASUREMENT?

The emphatic rejection of the axioms of representational measurement across New Zealand, Australia, Canada, and the United Kingdom is not the result of a single methodological error or historical accident. It reflects the formation and stabilization of a shared epistemic culture in health technology assessment that privileges arithmetic outputs over measurement validity. Once this culture took hold, it became self-reinforcing, reproducing itself through teaching, publication, and institutional practice while systematically excluding the theoretical constraints that would expose its core constructs as illegitimate.

A central driver of this rejection is path dependence. Early HTA frameworks adopted preference-based utilities and cost-effectiveness ratios at a time when formal measurement theory was largely absent from economics and health outcomes training. These constructs appeared numerically tractable and policy-friendly. Once embedded in national guidelines, journals, and postgraduate curricula, they acquired institutional authority independent of their scientific admissibility. Introducing representational measurement axioms at a later stage would not merely refine these methods; it would invalidate them. The axioms therefore represent an existential threat rather than a technical correction.

Another factor is the professional division of labor that shaped HTA’s intellectual foundations. Health economics evolved with a focus on optimization and decision rules, not on the conditions under which numbers legitimately represent empirical attributes. Measurement theory, particularly representational measurement and Rasch measurement, developed largely outside economics. As a result, HTA practitioners inherited a numerical toolkit without the theoretical apparatus required to police its use. The absence of measurement literacy was not perceived as a deficiency because the field never defined measurement as a prerequisite for arithmetic.

The appeal of multiattribute utility instruments further entrenched this stance. Instruments such as the EQ-5D and the Health Utilities Index offered a way to collapse complex health states into single numbers that could be manipulated algebraically. This convenience came at the cost of violating unidimensionality, invariance, and true-zero requirements. Accepting the axioms of representational measurement would force rejection of the premise that heterogeneous health attributes can be compressed into a single quantitative scale. Rather than confront this impossibility, HTA systems normalized it.

Simulation modeling provided an additional layer of insulation. Reference-case models and sensitivity analyses created the appearance of scientific rigor while bypassing the need for empirical testability. By shifting attention from measurement validity to internal model coherence, HTA systems could continue to generate precise numbers without confronting whether those numbers corresponded to measurable quantities. Measurement axioms were rendered irrelevant by a modeling culture that treated assumption-driven projections as evidence.

Institutional incentives also played a role. HTA agencies in these countries operate under strong pressure to deliver consistent, defensible recommendations within constrained timelines. Measurement-valid alternatives, such as single-attribute claims or Rasch-based latent trait measures, would require new protocols, new data collection, and new expertise. In contrast, the existing framework offers a standardized, internationally aligned process that can be defended procedurally even if it cannot be defended scientifically. The axioms of representational measurement disrupt this equilibrium.

Finally, there is a sociological dimension. Admitting that core HTA constructs violate fundamental measurement principles would imply that decades of decisions, publications, and training were built on indefensible foundations. Such an admission carries reputational and professional risk for individuals and institutions alike. The collective response has therefore been denial by omission. Measurement axioms are not debated; they are ignored.

New Zealand's alignment with Australia, Canada, and the United Kingdom reflects participation in this shared Anglophone HTA memeplex. The rejection of representational measurement is not explicit, but it is emphatic in practice. Arithmetic is performed confidently, models are elaborated, and numbers are treated as evidence. Measurement, however, is assumed rather than demonstrated. The result is a sophisticated evaluative apparatus that operates without the scientific constraints that would make its outputs meaningful.

CAN NEW ZEALAND TRANSITION TO REPRESENTATIONAL MEASUREMENT

The question of whether New Zealand can transition to representational measurement is not one of technical capacity, institutional intelligence, or analytical sophistication. It is a question of epistemic reordering. The existing HTA framework demonstrates that the country possesses substantial modeling expertise, well-developed administrative processes, and experienced evaluative committees. What it lacks is not competence, but a recognition of the logical sequence that makes quantitative claims possible: measurement must precede arithmetic.

At present, the New Zealand HTA system operates in reverse. Numerical manipulation is permitted first, and questions of measurement are either assumed away or treated as philosophical distractions. This ordering reflects inheritance rather than design. New Zealand did not independently construct its evaluative framework from foundational principles; it adopted a model already normalized internationally. In doing so, it also adopted the epistemic omissions embedded within that model. The consequence is a system capable of producing decisions but incapable of producing testable knowledge.

Transitioning to representational measurement therefore requires acknowledging a simple but destabilizing truth: many of the numerical objects currently used to justify pricing and access decisions are not measures at all. Utilities are ordinal preference expressions. QALYs are composite arithmetic artifacts. Simulation outputs are conditional projections. None possess the properties required to support multiplication, aggregation, or claims of magnitude. Recognizing this does not invalidate past decisions, but it does invalidate the belief that those decisions were evidence-based in the scientific sense.

The first requirement for transition is institutional recognition that measurement is not optional. Representational measurement theory specifies when numbers legitimately represent attributes and when they do not. These axioms are not negotiable conventions; they define the boundary between calculation and symbolism. Without explicit adoption of these axioms as gatekeeping criteria, no reform can occur. Transition cannot be achieved by modifying models, adjusting thresholds, or refining utility instruments. It requires redefining what counts as an admissible claim.

Once this recognition occurs, the pathway forward becomes conceptually straightforward. Therapy impact claims must be reformulated as single, unidimensional propositions. Each claim must specify the attribute being evaluated, the population to which it applies, and the timeframe over which it is expected to hold. Crucially, each claim must be classified as either manifest or latent. This distinction determines the only two admissible measurement standards available.

For manifest attributes, such as events, utilization, or time, linear ratio measurement is required. Counts must possess a true zero and invariant units. Where these conditions are met, arithmetic is legitimate. Where they are not, the claim must be rejected or reformulated. For latent attributes, such as symptom burden, functioning, or patient experience, Rasch measurement is mandatory. Without Rasch transformation, subjective responses remain ordinal and cannot support arithmetic under any circumstances.

This shift does not reduce analytical rigor. It increases it. It removes the illusion that complex modeling compensates for absent measurement and replaces it with claims that can be empirically evaluated, replicated, and potentially falsified. Importantly, it also narrows the evidenti burden. Instead of defending large composite models, manufacturers and evaluators focus on a limited number of clearly defined claims whose outcomes can be observed in real populations.

New Zealand is well positioned to make this transition precisely because its system is centralized and methodologically coherent. The same coherence that once facilitated adoption of the reference case can now facilitate reform. Transition does not require dismantling HTA institutions; it requires retraining them. Committees must be supported in understanding scale types, permissible arithmetic, and the distinction between measurement and scoring. Rasch measurement, often portrayed as technically intimidating, is in fact supported by mature software platforms that have been in continuous use for more than forty years. The barrier is not technology, but literacy.

The role of artificial intelligence is also transformative. Large language model diagnostics have already demonstrated that the absence of measurement axioms is not anecdotal but systematic. This changes the reform landscape fundamentally. The failure is now visible, reproducible, and comparable across jurisdictions. New Zealand is no longer isolated in confronting this issue; it is part of a global recognition that the emperor has no clothes.

Transition therefore does not represent an admission of failure. It represents a restoration of scientific legitimacy. A system grounded in representational measurement can generate claims that are provisional rather than performative, testable rather than rhetorical, and cumulative rather than circular. It replaces administrative closure with empirical accountability.

Whether New Zealand can transition depends not on resources, but on willingness. The tools exist. The theory is established. The diagnostic evidence is now unavoidable. What remains is a decision about the future identity of HTA itself. It can remain an administrative mechanism that produces numbers to justify decisions, or it can become a scientific enterprise that generates measurable claims about therapy impact. It cannot be both.

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