MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

NEW ZEALAND: THE ABSENCE OF MEASUREMENT IN ACADEMIC RESEARCH GROUPS FOR HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 36 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 purpose of this assessment is to evaluate the extent to which academic health technology assessment (HTA) research groups in New Zealand demonstrate understanding and application of the axioms of representational measurement. Using the canonical 24-item diagnostic framework, the analysis interrogates whether the academic knowledge base recognizes the fundamental requirement that measurement must precede arithmetic, and whether it distinguishes between admissible forms of measurement for manifest attributes and latent traits. The objective is not to evaluate individual publications or researchers, but to characterize the structural properties of the academic environment within which HTA methods are taught, reproduced, and legitimized.

Specifically, the assessment seeks to determine whether New Zealand academic centers operate within a measurement-valid framework capable of supporting falsifiable therapy impact claims, or whether they reproduce the global HTA reference-case memeplex in which ordinal constructs are treated as quantitative outcomes and arithmetic operations are applied without prior demonstration of scale properties. The analysis therefore focuses on the boundaries of admissible reasoning embedded in the academic corpus, rather than on methodological variation within those boundaries.

The results show that New Zealand academic HTA research groups exhibit the same epistemic structure observed across jurisdictions that have adopted the NICE reference-case framework. Canonical propositions that would enforce representational measurement collapse to low endorsement probabilities, while propositions that permit arithmetic without measurement are strongly reinforced. This pattern is not inconsistent or transitional; it is stable and internally coherent.

The diagnostic demonstrates that the academic knowledge base accepts, often implicitly, the legitimacy of QALYs, utility algorithms, summated questionnaire scores, and reference-case simulation models, while simultaneously failing to recognize the axioms that would be required to justify these practices. Latent attributes are routinely invoked but not measured. Rasch methodology is absent as a gatekeeping requirement. Falsification is reinterpreted as sensitivity analysis rather than empirical refutation. As a result, academic centers function not as sites of foundational scrutiny but as mechanisms of methodological replication, reinforcing the same numerical storytelling architecture that underpins national HTA decision making.

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 RESEARCH GROUP HTA KNOWLEDGE BASE

The knowledge base of academic health technology assessment in New Zealand can be characterized as a system organized around the production, refinement, and application of numerical outputs rather than the construction of scientifically admissible measures. Within this system, the presence of numbers is treated as sufficient evidence of quantification, and the distinction between ordering and measuring is rarely acknowledged as analytically decisive. Measurement is assumed rather than demonstrated.

At the center of this knowledge base lies the normalization of composite and summated outcomes. Subjective responses collected through questionnaires, preference instruments, and health-state descriptions are routinely aggregated into total scores or utility weights that are subsequently treated as continuous variables. These outputs are interpreted as magnitudes of effect, improvements in health, or differences between interventions, despite the absence of demonstrated equal intervals, invariance, or meaningful zero points. Statistical properties such as reliability, responsiveness, and internal consistency are treated as substitutes for measurement rather than as preliminary descriptive diagnostics.

Latent attributes play a central rhetorical role within the academic literature. Concepts such as quality of life, wellbeing, health status, and burden are repeatedly described as quantities subject to change and comparison. Yet these attributes are not operationalized through measurement models capable of producing invariant units. The notion of latent trait possession is largely absent. Instead, latent constructs are treated as if they were directly observable entities whose magnitude can be inferred from summed responses. This approach collapses the distinction between classification and measurement, allowing ordinal data to masquerade as quantitative outcomes.

The exclusion of Rasch measurement is decisive in defining the boundaries of this knowledge base. Rasch methodology, which uniquely satisfies the requirements for constructing linear logit ratio measures for latent traits, is not integrated as a necessary condition for evaluable claims. Its absence allows the continued use of multidimensional instruments, domain-weighted composites, and preference-based algorithms without confronting their failure to meet unidimensional measurement requirements. Without Rasch transformation, subjective data remain ordinal regardless of subsequent statistical manipulation, yet this structural limitation is not treated as analytically consequential.

Arithmetic operations are therefore applied permissively. Means, differences, regressions, and modeled projections are performed without reference to permissible transformations. Change scores are interpreted as quantitative improvement. Between-group comparisons are interpreted as magnitude effects. Simulation outputs are treated as evidence rather than as conditional narratives. In this environment, analytical complexity substitutes for measurement validity.

The knowledge base also aligns closely with international HTA conventions. The reference-case framework is accepted as methodological best practice, not because it satisfies scientific axioms, but because it provides procedural closure. Models deliver numbers that appear precise, reproducible, and decision-relevant. This appearance of rigor allows academic work to interface seamlessly with policy processes while avoiding the unresolved question of whether the outputs represent measurable quantities.

What defines the New Zealand academic HTA knowledge base most clearly is not explicit denial of measurement theory, but patterned silence. Representational measurement theory is not debated; it is absent. Stevens' scale typology is rarely operationalized. The requirement that measurement precede arithmetic is not enforced as a gatekeeping criterion. This absence creates a permissive epistemic environment in which numerical storytelling can flourish without challenge.

As a result, academic centers function primarily as transmitters of an inherited framework rather than as evaluators of its legitimacy. The system reproduces itself through teaching, publication, and professional socialization, ensuring continuity while insulating its core assumptions from scrutiny. The outcome is a stable yet scientifically fragile knowledge base that generates numbers efficiently while remaining detached from the foundational conditions required for those numbers to mean anything at all.

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 RESEARCH GROUPS

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 RESEARCH GROUPS

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.25	-1.10
MEASURES MUST BE UNIDIMENSIONAL	1	0.30	-0.85
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.20	-1.40
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75

RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.85	+1.75
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
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	0.20	-1.40
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.20	-1.40
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	0.25	-1.10
QALYS CAN BE AGGREGATED	0	0.90	+2.20
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.65	+0.60
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.80	+1.40
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.85	+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	0.30	-0.85

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

NEW ZEALAND ACADEMIC HTA RESEARCH GROUPS: MEASUREMENT WITHOUT MEASUREMENT

The Table 1 diagnostic profile for New Zealand academic research centers engaged in health technology assessment reveals a structure that is not anomalous, not transitional, and not confused. It is internally coherent. What it lacks is measurement. The probabilities and logits do not describe a contested intellectual environment. They describe a settled one. Across the canonical statements, the pattern is stable and unmistakable: propositions that would enforce representational measurement collapse toward the floor, while propositions that permit arithmetic without measurement cluster toward the ceiling. This is not ignorance in the everyday sense. It is institutionalized absence.

The defining feature of the New Zealand academic HTA knowledge base is that it treats numerical form as sufficient evidence of quantification. Once an outcome is expressed as a number, the question of what that number represents disappears. Measurement is presumed, never demonstrated. Arithmetic becomes permissible by convention rather than by proof. This is visible immediately in the treatment of scale types. The proposition that interval measures lack a true zero is weakly endorsed at $p = 0.25$. The proposition that multiplication requires ratio measurement sits even lower. These are not controversial claims in measurement theory; they are definitional. Their rejection indicates that scale properties are not functioning as gatekeeping criteria within academic practice. Instead, scale type is treated as descriptive language rather than as a logical constraint.

This explains why the QALY is endorsed so strongly. The claim that the QALY is a ratio measure sits near the ceiling at $p = 0.90$. The claim that QALYs can be aggregated sits equally high. The claim that EQ-5D algorithms produce interval measurement is also reinforced at near-ceiling levels. These endorsements are not empirical judgments. They are functional necessities. Without them, the entire analytical enterprise collapses. The academic literature does not arrive at these conclusions through demonstration. It inherits them. Utility algorithms are accepted because they are used. QALYs are treated as ratio measures because policy requires multiplication. Negative utilities are normalized because models permit them. Each assumption exists because the system cannot function without it.

The diagnostic captures this circularity precisely. When the proposition "measurement precedes arithmetic" falls to $p = 0.20$, the inversion becomes explicit. Arithmetic is permitted first. Measurement is assumed afterward. This inversion is the epistemic signature of the HTA memeplex. Nowhere is this more evident than in the treatment of latent attributes. New Zealand academic centers routinely invoke constructs such as quality of life, wellbeing, health status, burden, and functioning. These are explicitly latent attributes. They cannot be observed directly. They require construction through a measurement model. Yet the proposition that Rasch transformation is required for latent traits collapses to p = 0.10 with a logit of −2.20. This is not ambivalence. It is exclusion. Rasch measurement is not debated within the academic knowledge base. It is absent. The concept of latent trait possession barely registers. Instead, latent attributes are treated as if they were directly observable quantities. Questionnaire responses are summed. Domain scores are averaged. Change scores are computed. These numerical artifacts are then analyzed using regression, mixed models, and sensitivity analysis, all without ever establishing whether the numbers possess equal intervals or invariant meaning.

The proposition that summated Likert scores create ratio measures sits at p = 0.90. This single endorsement explains the entire structure. Once summation is believed to manufacture quantity, every downstream operation becomes legitimate. Means can be calculated. Differences interpreted. Effect sizes reported. Models populated. No further epistemic justification is required. This belief is not corrected by statistical sophistication. In fact, sophistication reinforces it. Advanced modeling techniques obscure rather than resolve measurement failure. Precision substitutes for validity. Confidence intervals create the appearance of rigor while masking the absence of units.The diagnostic also reveals how falsification has been redefined. The statement that non-falsifiable claims should be rejected receives moderate endorsement. In isolation, this suggests commitment to scientific norms. Yet the statement that reference-case simulation generates falsifiable claims is endorsed strongly. This contradiction is resolved not by logic but by semantic drift. Falsification is no longer understood as exposure to empirical refutation. It is reinterpreted as robustness testing within a model. Sensitivity analysis becomes a surrogate for falsification. If a model's outputs remain stable under parameter variation, the claim is treated as credible. But nothing in such exercises exposes the claim to the world. The outcome cannot be wrong in the Popperian sense. It can only be different under different assumptions. This is not science. It is scenario narration.

New Zealand academic centers have adopted this reinterpretation fully. It allows the system to speak the language of science while avoiding its obligations. Claims appear empirical while remaining insulated from disconfirmation. This epistemic posture explains the extraordinary stability of the knowledge base. The probabilities vary slightly across assessments, but the logits remain clustered. This is precisely what one would expect when beliefs are structurally reinforced rather than individually reasoned. The knowledge base does not depend on particular authors or institutions. It is reproduced through training, journals, guidelines, and professional norms.

Students entering academic programs are taught methods, not axioms. They learn how to construct models, not when modeling is permissible. They learn how to manipulate utilities, not what utilities are. Stevens' typology of measurement scales is rarely taught. Representational measurement theory is absent. Rasch is treated as psychometric history rather than as the only legitimate framework for latent trait measurement. As a result, graduates return to ministries, consultancies, and HTA agencies with procedural competence but no epistemic defenses. They can implement the reference case flawlessly while lacking the conceptual tools to ask whether it makes sense. This explains why New Zealand did not resist the NICE framework. There was nothing in the academic knowledge base capable of resisting it. The framework arrived already numerically articulated, already institutionally legitimated, already encoded in journals and textbooks. To challenge it would have required concepts that were not taught. The result is a perfect transmission environment. The reference case does not need persuasion. It requires only acquiescence. And acquiescence is guaranteed when measurement theory is absent. The canonical logits reveal something even more troubling. The strongest endorsements cluster precisely around the propositions that enable closure. QALYs as ratio measures. Aggregation across persons. Simulation as evidence. These beliefs are not random errors. They are administrative enablers. They allow decisions to be made.

Academic centers function less as traditional sites of discovery than as suppliers of justifications. They do not generate falsifiable claims. They generate defensible narratives. Their outputs are optimized not for truth, but for usability. They have, apparently, abandoned any notion of therapy impact analysis that is consistent with the evolution of objective knowledge. This is why the title “academic research center” becomes misleading. Research implies uncertainty, testing, and revision. What exists instead is methodological reproduction. Once a framework is accepted, the role of academia becomes refinement, extension, and application, never interrogation.

The absence of Rasch measurement is decisive here. Rasch would force confrontation with unidimensionality, item invariance, and scale construction. It would expose the impossibility of multiattribute quality of life indices. It would collapse the utility framework. Its exclusion is therefore not accidental. It is structurally necessary. What emerges from the diagnostic is not incompetence, nor bad faith. It is a closed epistemic economy. Academic centers receive legitimacy by conforming to international norms. Journals reward compliance. Funding favors alignment. Deviations are seen as impractical rather than insightful.

Within such an environment, asking whether arithmetic is permitted becomes socially irrational even if logically necessary. The most striking implication is that New Zealand’s academic HTA centers do not merely fail to correct the system. They stabilize it. They provide the intellectual continuity that allows agencies to believe the framework is scientifically grounded. They serve as validators rather than critics. This is why academic centers are the critical failure point. Agencies can be forgiven for seeking closure. Politicians can be forgiven for demanding decisions. But universities are meant to ask whether the numbers mean anything at all. In New Zealand they do not. Yet this also identifies the path forward. If transition to representational measurement is to occur, it must begin in academic training. Without measurement literacy, no reform can hold. Committees cannot enforce what scholars cannot explain. Agencies cannot require what universities do not teach.

CAN ACADEMIC RESEARCH CENTERS IN NEW ZEALAND EMBRACE REPRESENTATIONAL MEASUREMENT

Whether academic research centers in New Zealand can embrace representational measurement is not a question of technical capacity. It is a question of intellectual realignment. The axioms of representational measurement are neither new nor obscure. They have been available for over seventy years, formalized through Stevens’ scale typology and extended through the development of Rasch measurement for latent traits. What has been missing is not access to knowledge, but recognition that measurement is not optional. It is a precondition for quantitative inference.

At present, academic HTA centers operate within a framework in which numbers are treated as evidence by default. This environment rewards analytic fluency within the reference-case architecture rather than scrutiny of its foundations. Students are trained to build models, generate utilities, and manipulate QALYs without ever being asked what kind of numbers these are. As a result, the academic system does not merely fail to teach representational measurement; it implicitly teaches that such questions are unnecessary. This creates a structural barrier to reform that is cultural rather than methodological.

Yet nothing intrinsic to New Zealand academia prevents transition. The universities possess statistical expertise, methodological sophistication, and research infrastructure equal to any international center. What they lack is a gatekeeping doctrine that insists measurement must precede arithmetic. Once that ordering is restored, much of the existing analytical machinery becomes immediately questionable. Ordinal utilities can no longer be multiplied by time. Summated questionnaire scores can no longer be treated as magnitudes. Simulation outputs can no longer be described as evidence unless their dependent variables meet measurement standards. This is uncomfortable, but not impossible.

The most significant obstacle is professional inheritance. Academic HTA centers did not design the reference-case framework; they inherited it. Careers have been built within it. Journals expect it. Funding agencies assume it. International alignment depends upon it. Embracing representational measurement therefore appears, incorrectly, as an act of rebellion rather than correction. In reality, it is a return to scientific normality. Measurement discipline is not innovation; it is restoration.

Transition becomes possible once the distinction between manifest and latent attributes is made explicit. For manifest claims—events, time, utilization, counts—the requirement is straightforward: linear ratio measurement with a true zero. For latent attributes, symptom burden, functioning, need fulfillment, the requirement is equally clear: Rasch logit ratio measurement. These two forms exhaust the admissible options. Everything else is descriptive scoring, not measurement. Once academic centers accept this classification, the reform pathway becomes coherent rather than threatening.

Critically, embracing representational measurement does not mean abandoning relevance to policy. It means abandoning false precision. Health systems do not need composite value constructs to make decisions; they need credible, evaluable claims. Single-attribute claims measured properly are far more informative than multidimensional indices that cannot be falsified. Academic centers are uniquely positioned to lead this transition because they train the next generation of analysts. If universities continue to teach arithmetic without measurement, the system will reproduce failure indefinitely.

The emergence of AI-based LLM diagnostics has altered the landscape decisively. For the first time, belief structures can be examined empirically across entire knowledge bases. The absence of measurement axioms is no longer speculative; it is demonstrable. This creates both exposure and opportunity. Academic centers can no longer plausibly claim ignorance. But they can claim leadership by acknowledging the failure and initiating reform.

Embracing representational measurement therefore requires institutional courage, not methodological reinvention. It requires academic programs to reintroduce foundational measurement theory, to teach Rasch not as a niche psychometric technique but as the only defensible approach to latent trait quantification, and to treat measurement validity as a gatekeeping requirement rather than a post hoc justification. If New Zealand academic research centers are willing to make that shift, they can move from being transmission nodes of the reference-case memeplex to becoming catalysts for its replacement.

The choice is stark but constructive. Continue teaching numerical storytelling, or restore the conditions under which quantitative science is possible. The tools already exist. What remains uncertain is the willingness to use them.

CAN PHARMAC RESPOND

Whether PHARMAC can respond to the measurement critique is not a question of institutional intelligence or administrative competence. PHARMAC is widely regarded as one of the most disciplined purchasing agencies in the world. Its challenge lies elsewhere. It has inherited an evaluative framework whose numerical outputs appear authoritative while lacking the properties required for scientific measurement. The question, therefore, is not whether PHARMAC can operate the existing system efficiently, but whether it can recognize that efficiency within a false measurement framework cannot produce evaluable knowledge.

PHARMAC's current assessment architecture mirrors the international HTA memeplex. It relies on modeled claims, preference-weighted health states, QALYs, and long-horizon simulations to support pricing and access decisions. These tools provide administrative closure, but they do not permit falsification. Once a model is accepted and a decision is made, there is no empirical mechanism to test whether the claim was correct. This is not a technical oversight; it is a structural consequence of arithmetic being applied in the absence of demonstrable measurement.

Responding to the critique therefore requires PHARMAC to confront a difficult realization. The framework it uses was not designed to support normal science. It was designed to manage uncertainty through consensus and convention rather than through provisional, testable claims. This explains its durability. It also explains its scientific vulnerability. A system that cannot generate claims capable of being wrong cannot learn from experience, no matter how sophisticated its modeling appears.

The obstacle PHARMAC faces is not that the critique is complex, but that it challenges the starting point. Representational measurement requires that numerical claims be anchored in scale properties before arithmetic is permitted. For manifest attributes such as events, utilization, or time, this means linear ratio measures with a true zero. For latent attributes such as functioning or patient experience, it means Rasch logit ratio measurement. PHARMAC's present framework does not enforce either condition. As a result, it evaluates models rather than claims.

Yet PHARMAC is better positioned than most agencies to respond. Unlike many HTA bodies, it already emphasizes budget impact, real-world consequences, and accountability to a defined population. These priorities align naturally with a single-claim measurement framework. A transition away from reference-case modeling toward explicitly defined, unidimensional claims would not weaken PHARMAC's mandate. It would strengthen it by enabling post-listing evaluation, replication, and correction.

What prevents response is not institutional incapacity, but epistemic inertia. For decades, health technology assessment has taught agencies that measurement questions were settled long ago. Utilities were assumed to be interval. QALYs were assumed to be ratio. Simulation was assumed to approximate evidence. These assumptions were rarely challenged because few within the system were trained to challenge them. PHARMAC did not choose this ignorance; it inherited it.

The emergence of LLM-based diagnostics changes that inheritance. The absence of measurement axioms across national and institutional knowledge bases can now be demonstrated rather than asserted. This creates a moment of unavoidable clarity. PHARMAC can no longer treat the framework as merely "what everyone does." The emperor has no clothes, and that fact is now visible at scale.

Responding does not require PHARMAC to abandon decision making, nor to suspend access while waiting for perfect data. It requires a shift in what counts as admissible evidence. Instead of accepting composite modeled outputs, PHARMAC can require manufacturers to present a limited number of single, measurable claims, each supported by an agreed protocol and subject to evaluation over time. This restores falsification, learning, and accountability.

The real question, therefore, is not whether PHARMAC can respond. It is whether it is willing to replace administrative certainty with scientific humility. Measurement forces provisional claims. It accepts that claims may fail. It rejects the comfort of final numbers. But it is the only path by which evidence can evolve.

PHARMAC has long prided itself on independence and rigor. Embracing representational measurement would not diminish that reputation. It would redefine it.

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

.

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