

MAIMON RESEARCH LLC
**ARTIFICIAL INTELLIGENCE LARGE LANGUAGE
MODEL INTERROGATION**



**REPRESENTATIONAL MEASUREMENT FAILURE IN
HEALTH TECHNOLOGY ASSESSMENT**

**UNITED STATES: PARADIGM FAILURE AND
MEASUREMENT INVERSION IN HEALTH
TECHNOLOGY ASSESSMENT**

**Paul C Langley PhD Adjunct Professor, College of Pharmacy, University of
Minnesota, Minneapolis, MN**

LOGIT WORKING PAPER No 657 JUNE 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

Health technology assessment (HTA) has, for more than four decades, relied upon a reference-case analytical framework centered on utilities, quality-adjusted life years (QALYs), cost-effectiveness ratios, and simulation modelling. These methods have become the accepted basis for evaluating therapeutic value in the United States and have influenced reimbursement systems worldwide. Their widespread acceptance, however, has rarely been accompanied by a critical examination of whether they satisfy the fundamental requirements of quantitative science. This paper addresses that question through an interrogation of the knowledge bases of seven leading United States HTA research centres using a large language model (LLM) framework grounded in the principles of representational measurement.

The interrogation employs twenty-four canonical statements that distinguish between propositions that are true and false under measurement theory. Categorical endorsement probabilities and normalized logits are used to estimate the extent to which the HTA knowledge base recognizes the principles governing lawful measurement, including scale theory, ratio measurement, unidimensionality, dimensional homogeneity, admissible arithmetic, representational measurement, Rasch measurement, falsifiability, and latent trait measurement. The results reveal a remarkably consistent pattern across all seven research centers. Foundational principles of measurement science receive uniformly weak endorsement, while propositions underpinning utilities, QALYs, and reference-case simulation models receive consistently strong endorsement. The resulting profile is interpreted as compelling evidence of measurement inversion, in which arithmetic is allowed to precede measurement.

The paper argues that this pattern cannot be understood in isolation but reflects the educational phenomenon of curriculum inversion, whereby students are trained to construct utilities, perform economic evaluations, and build simulation models without first acquiring the conceptual foundations necessary to determine whether those analyses are scientifically legitimate. Curriculum inversion reproduces measurement inversion across successive generations of researchers, educators, and policy makers, thereby sustaining a paradigm detached from the standards of normal science. The paper concludes that incremental refinement of the existing framework is insufficient. Instead, it proposes a transition to a measurement-based HTA paradigm founded on representational measurement, linear ratio measures for manifest attributes, and Rasch logit ratio measures for latent attributes. Together, these provide the only scientifically defensible basis for evaluable, replicable, and falsifiable claims regarding therapeutic impact and value.

INTRODUCTION

The extent to which the contemporary analytical framework of health technology assessment (HTA), the reference-case simulation model, can continue to claim scientific legitimacy is a critical question for its future survival as a framework for evaluating therapeutic value. For more than four decades, the reference case has dominated HTA in the United States and has exerted a profound influence on reimbursement decision making throughout the world. Utilities, quality-adjusted life years (QALYs), cost-effectiveness ratios, and simulation models have become accepted features

of the HTA landscape and are widely regarded as the defining components of modern therapeutic evaluation.

Yet acceptance is not validation. The central question is whether the architects of the reference-case framework, and those institutions responsible for its continuing development and promotion, recognized the fundamental requirement that measurement must precede arithmetic. Before quantities can be multiplied, divided, aggregated, averaged, or incorporated into simulation models, their measurement status must first be established. This principle is not unique to HTA. It is a foundational requirement of quantitative science and is embodied in the axioms of representational measurement, the theory of measurement scales, and the conditions governing admissible arithmetic.

The evidence accumulated to date leaves little room for ambiguity. Across a growing series of interrogations of HTA agencies, academic centers, professional organizations, journals, and educational programs, the dominant pattern has been one of measurement inversion. Rather than establishing measurement properties before undertaking arithmetic operations, the contemporary HTA framework proceeds in the opposite direction. Arithmetic is accepted as legitimate while measurement is assumed. Utility scores are treated as though they possess ratio properties. QALYs are constructed without demonstrating dimensional homogeneity. Simulation models manipulate quantities whose measurement status remains unresolved. The result is a framework built upon assumptions regarding measurement rather than measurement itself.

RESEARCH CENTERS

The research centers that are the focus for this assessment of measurement inversion are the Schaeffer Center for Health Policy and Economics at the University of Southern California ¹, the Harvard Center for Health Decision Science ², the CHOICE Institute at the University of Washington ³, the Duke-Margolis Institute for Health Policy ⁴, the Leonard Davis Institute of Health Economics at the University of Pennsylvania ⁵, the Bloomberg School of Public Health at Johns Hopkins University ⁶ and the Stanford Health Policy Center ⁷.

If these centers demonstrate limited recognition of the standards governing measurement, then the implications extend far beyond individual institutions. They raise fundamental questions regarding the scientific foundations of the reference-case paradigm itself. The objective of this assessment is therefore not merely to examine the knowledge profiles of these research groups, but to determine whether the intellectual foundations of US HTA remain consistent with the requirements of quantitative science.

STANDARDS FOR MEASUREMENT

The starting point for any scientific discipline that seeks to make quantitative claims is measurement. Before quantities can be manipulated mathematically, it must first be demonstrated that they possess the properties necessary to support the proposed arithmetic operations. This principle is fundamental to both the physical and social sciences. Measurement precedes arithmetic. Quantitative claims are valid only when the quantities involved satisfy the requirements

of measurement. If these requirements are absent, arithmetic operations may still be performed, but the resulting outputs have no scientific standing as measures.

The importance of this principle is reflected in the theory of measurement scales⁸. Not all numerical assignments possess the same properties. Nominal scales classify. Ordinal scales rank. Interval scales support differences between values. Ratio scales alone support the full range of arithmetic operations because they possess a true zero and permit proportional comparisons. Consequently, the admissibility of arithmetic depends upon scale type. Addition and subtraction require at least interval properties. Multiplication and division require ratio properties. This is not a matter of convention. It is a requirement imposed by the structure of measurement itself.

The central importance of ratio measurement follows directly from these considerations. Any claim involving multiplication, division, proportional comparison, growth rates, averages of ratios, or cost-effectiveness ratios requires quantities that possess ratio properties. If ratio measurement has not been demonstrated, these operations are inadmissible. Numerical manipulation cannot create measurement properties that are absent from the underlying scale. Arithmetic cannot substitute for measurement.

These requirements are formalized in the axioms of representational measurement⁹. Representational measurement provides the scientific framework that links empirical observations to numerical representations. Its purpose is to ensure that numerical assignments preserve the structure of the attribute being measured. Only when this correspondence is demonstrated can arithmetic operations be regarded as meaningful. The axioms of representational measurement therefore establish the conditions under which quantitative claims can be considered scientifically legitimate.

Among the most important of these requirements is unidimensionality. Measurement requires that an attribute represent a single dimension. If multiple attributes are combined into a composite score, numerical aggregation may be possible, but measurement has not necessarily occurred. Without unidimensionality there is no assurance that a numerical value represents a coherent quantity. The distinction between aggregation and measurement is therefore fundamental. Numbers can always be combined. Measures cannot be assumed.

Equally important is the distinction between manifest and latent attributes. Manifest attributes are directly observable and, where appropriately specified, support linear ratio measurement. Latent attributes are not directly observable and require a measurement model capable of estimating possession of the attribute. In the latter case, the required measure is the Rasch logit ratio scale¹⁰. These two forms of ratio measurement, linear ratio measurement for manifest attributes and Rasch logit ratio measurement for latent attributes, provide the only scientifically defensible basis for quantitative claims regarding therapy impact.

Taken together, these principles establish a clear standard. Measurement must precede arithmetic. Scale properties determine admissible operations. Ratio measurement is required wherever proportional comparisons or multiplication are involved. Unidimensionality must be demonstrated before measurement can be claimed. Representational measurement provides the governing scientific framework. Any discipline seeking to generate quantitative claims must satisfy these

requirements. Without them, numerical outputs remain constructions rather than measures, and quantitative claims become matters of assumption rather than science.

UNDERSTANDING MEASUREMENT INVERSION

Measurement inversion occurs when arithmetic is allowed to precede measurement. Instead of first establishing whether an attribute satisfies the requirements for measurement and then determining which mathematical operations are admissible, the sequence is reversed. Numerical operations are performed first and the resulting outputs are subsequently treated as though they possess measurement status. In effect, measurement is assumed rather than demonstrated.

This reversal lies at the heart of the contemporary HTA paradigm. In normal scientific practice, measurement serves as the foundation for quantitative claims. Before quantities can be added, multiplied, divided, averaged, aggregated, or incorporated into statistical and simulation models, their scale properties must be established. The distinction between nominal, ordinal, interval, and ratio scales determines which arithmetic operations are lawful. Arithmetic follows measurement because arithmetic derives its legitimacy from measurement.

Measurement inversion abandons this principle. Quantities are treated as measures because they are expressed numerically. Utility scores are assumed to possess the properties required for multiplication. QALYs are assumed to be valid quantitative outcomes because they result from arithmetic operations. Simulation models are assumed to generate evidence because they manipulate numerical inputs. The scientific burden of demonstrating measurement is quietly displaced by the assumption that numerical construction itself creates legitimacy.

The consequences are far-reaching. Once arithmetic is detached from measurement, the distinction between numbers and measures disappears. Ordinal responses become treated as though they were interval or ratio quantities. Composite indices are interpreted as measures. Utility algorithms are assumed to create measurable entities. Mathematical sophistication becomes confused with scientific validity. Increasingly complex calculations then conceal rather than resolve the absence of measurement foundations.

The problem is particularly evident in the treatment of utilities and QALYs. Utility scores are derived from health-state valuations and subsequently multiplied by time to generate QALYs. Yet multiplication requires ratio measurement. Unless both components possess ratio properties, the operation is inadmissible. The issue is not whether the resulting number is useful, convenient, or widely accepted. The issue is whether the arithmetic itself is lawful. Measurement inversion avoids this question by assuming that measurement has already been achieved.

The same logic applies to simulation modelling. Simulation models do not create measurement. They merely manipulate quantities supplied to them. If those quantities lack measurement status, the model cannot remedy the defect. More sophisticated modelling simply extends the underlying assumptions across larger populations, longer time horizons, and increasingly elaborate scenarios. The result is not stronger evidence but a more elaborate numerical construction.

The significance of measurement inversion therefore extends beyond individual methods. It represents a reversal of the scientific process itself. Rather than allowing measurement principles to govern arithmetic, arithmetic becomes the source of legitimacy. Quantitative claims are accepted because calculations can be performed, not because measurement has been demonstrated. The result is a framework in which numerical outputs acquire authority while the conditions necessary to support that authority remain largely unexamined.

This is why measurement inversion is so important. It provides a unifying explanation for the persistence of utilities, QALYs, cost-effectiveness ratios, and reference-case simulation models despite their failure to satisfy the requirements of representational measurement. It identifies the point at which HTA departed from the standards of quantitative science. More importantly, it explains why reconstruction must begin with measurement. Until measurement is restored to its proper place as the prerequisite for arithmetic, the scientific legitimacy of quantitative claims regarding therapy impact will remain fundamentally compromised.

INTERROGATING THE KNOWLEDGE BASE

A large language model (LLM) is an artificial intelligence system trained on extensive collections of text to identify patterns, relationships, concepts, and structures within language. Rather than retrieving information from a fixed database, an LLM generates responses by estimating the most probable continuation of text based upon the knowledge patterns represented within its training corpus together with any additional information available at the time of interrogation. These additional sources may include publicly available web content, user-supplied documents, and accumulated contextual memory derived from prior interactions and analyses. Consequently, the effective knowledge base available to an interrogation extends beyond the model's original training cut-off and reflects the broader information environment to which the model has access.

In this analysis, “interrogation” refers not to discovering what an LLM believes. An LLM possesses neither beliefs nor opinions. Rather, interrogation refers to probing the content of a corpus-defined knowledge space. The objective is to characterize the concepts, assumptions, relationships, and analytical frameworks that are embedded within a defined body of literature and professional discourse. In the present study, the interrogation also draws upon approximately twelve months of accumulated HTA-related analyses undertaken using the same framework. This provides additional contextual information regarding the conceptual structure of the HTA literature and its treatment of measurement-related issues.

The corpus is defined before interrogation. Depending upon the objective of the study, it may consist of a journal, a national HTA agency, a professional organization, a methodological framework, a curriculum environment, a research Centre, or a collection of policy documents. Once the boundaries of the corpus have been established, the LLM is used to estimate its conceptual footprint. This approach allows identification of concepts that are strongly reinforced, weakly represented, neglected, misunderstood, or systematically excluded.

In the present assessment, the objective is precise: to determine the extent to which a given HTA knowledge base recognizes and reinforces the foundational principles of representational measurement theory. The core proposition under investigation is that measurement precedes

arithmetic. No quantity should be treated as a number or subjected to mathematical operations unless the requirements of measurement have first been satisfied. These requirements include recognition of scale-type distinctions, unidimensionality, invariance, admissible arithmetic, dimensional homogeneity, and the principle that ordinal observations cannot be transformed into interval or ratio measures except under Rasch measurement rules.

The HTA knowledge space is defined pragmatically and operationally. Depending upon the institution or jurisdiction under examination, the corpus may include published HTA guidelines, agency decision frameworks, cost-effectiveness reference cases, academic journals and textbooks, modelling templates, technical reports, task-force recommendations, teaching materials, methodological articles, institutional white papers, and associated research publications. Collectively, these sources form the epistemic environment within which HTA practitioners develop their analytical frameworks, interpret evidence, and justify evaluative practices. The boundary of interrogation is therefore not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning.

The interrogation process relies on categorical endorsement probabilities. Endorsement does not imply that a statement appears verbatim within the corpus. Rather, it reflects the extent to which the concepts embodied in a statement are supported, reinforced, or implied by the knowledge base as a whole. A high endorsement probability indicates that a proposition is broadly consistent with the dominant conceptual structure of the corpus. A low endorsement probability indicates that the proposition is weakly represented, rarely articulated, absent, or contradicted by prevailing assumptions.

This distinction is crucial. The purpose is not to identify isolated references to representational measurement, ratio measurement, unidimensionality, Rasch measurement, or falsification. The purpose is to determine whether these concepts occupy a meaningful place within the intellectual environment. A single article mentioning unidimensionality does not imply that unidimensionality is an organizing principle of a discipline. Conversely, repeated reliance on utilities, QALYs, and simulation models may indicate strong endorsement of assumptions even when those assumptions are not explicitly stated.

The resulting endorsement profile therefore provides a measure of the conceptual structure of the knowledge base. Statements receiving consistently high endorsement identify concepts that are reinforced within the HTA environment. Statements receiving consistently low endorsement identify concepts that are neglected, marginalized, or absent. When propositions grounded in measurement science receive low endorsement while propositions embedded in the reference-case paradigm receive high endorsement, the resulting pattern is interpreted as evidence of measurement inversion. The interrogation therefore evaluates not individual opinions but the intellectual architecture of the HTA knowledge base itself.

CATEGORICAL ENDORSEMENT 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 a 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 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

RESULTS: A CONSENSUS PATTERN OF MEASUREMENT INVERSION

The results for the seven US HTA academic knowledge base interrogations are summarized in Table 1. For each of the 24 canonical statements, the table identifies whether the statement is true or false under the axioms of representational measurement, together with the observed range of categorical probabilities and the corresponding range of normalized logits across the five domains. As in the Australian and Canadian analyses, the purpose of presenting ranges is to make explicit both the level and the consistency of endorsement. The question is not whether individual values differ, but whether the pattern of endorsement varies in direction or structure across knowledge bases.

TABLE 1: SUMMARY OF STATEMENT RESPONSES: SEVEN US RESEARCH CENTRE KNOWLEDGE BASE DOMAINS

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY RANGE	NORMALIZED LOGIT (IN RANGE +/- 2.50) RANGE
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.60
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.80 to 0.85	+1.40 to +1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.80 to 0.85	+1.40 to +1.75
THE QALY IS A RATIO MEASURE	0	0.85 to 0.90	+1.75 to +2.20
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.85	+1.75
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85 to 0.90	+1.75 to +2.20
CLAIMS FOR COST-EFFECTIVENESS FAIL THE	1	0.15 to 0.20	-1.40 0 -1.75

AXIOMS OF REPRESENTATIONAL MEASUREMENT			
QALYS CAN BE AGGREGATED	0	0.85 to 0.90	+1.75 to +2.20
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60 to 0.65	+0.70 to +0.85
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85 to 0.90	+1.75 to +2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.60 to 0.65	+0.70 to +0.85
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.25 to 0.35	-1.90 to -1.25
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.90
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

Note: If there is not a range then the categorical probabilities are identical for all 7 research centers

The results summarized in Table 1 demonstrate a strikingly consistent pattern across the seven United States HTA academic research-center knowledge bases. The importance of the table lies not simply in the probabilities assigned to individual statements, but in the direction, stability and structure of the endorsement profile. Across all seven domains, propositions required by representational measurement receive low endorsement probabilities, while propositions embedded in the contemporary reference-case paradigm receive moderate to high endorsement. This is the signature of measurement inversion.

The most important finding is the collapse of endorsement for the foundational propositions of measurement science. The statement that measurement precedes arithmetic receives a categorical probability of only 0.10. The statement that multiplication requires a ratio measure also receives 0.10. The proposition that meeting the axioms of representational measurement is required for arithmetic similarly receives 0.10. These are not marginal or specialist propositions. They are the conditions that determine whether quantitative claims can be made at all. If measurement does not precede arithmetic, then there is no principled basis for deciding whether addition, multiplication, division, aggregation or ratio construction is admissible.

The same pattern is evident for unidimensionality and scale status. Measures must be unidimensional receives a probability of only 0.15. Interval measures lack a true zero receives only 0.20. These results indicate weak recognition of two basic requirements for measurement. Without unidimensionality, a numerical output may represent aggregation, scoring or indexing, but not measurement of a single attribute. Without recognition of the true-zero requirement, the distinction between interval and ratio scales collapses. This matters because the reference-case framework depends upon ratio operations. Utility scores are multiplied by time, QALYs are aggregated, and cost-effectiveness ratios are constructed. Yet the principles governing those operations receive little endorsement.

The results for Rasch measurement are even more decisive. The proposition that there are only two classes of measurement relevant to therapy assessment, linear ratio and Rasch logit ratio, receives a probability of only 0.05. The proposition that subjective responses can be transformed into measurement only through Rasch rules also receives 0.05. The proposition that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits receives 0.05. The proposition that Rasch rules embody the requirements of representational measurement receives 0.05. These are the lowest probabilities in the table, and their uniformity is important. They suggest that the conceptual framework required for latent attribute measurement is almost entirely absent from the US academic HTA knowledge base.

This absence is especially serious because contemporary HTA relies heavily on latent attributes. Pain, fatigue, functioning, quality of life, treatment satisfaction, psychological wellbeing and patient experience are not directly observable. They require a measurement model capable of estimating possession of the latent attribute. Without such a model, ordinal responses remain ordinal responses. Summation, weighting, transformation or incorporation into utility algorithms does not create measurement. The low probabilities for Rasch statements therefore indicate not merely the absence of a particular technique, but the absence of the measurement framework required to support quantitative claims for latent outcomes.

In contrast, propositions associated with the reference-case paradigm receive consistently high endorsement. Statements involving time trade-off preferences, EQ-5D-3L preference algorithms, QALYs, summated subjective responses, Likert score summation, dimensional homogeneity, QALY aggregation and reference-case simulation receive probabilities ranging largely from 0.80 to 0.90. The knowledge bases behave as though these constructs are legitimate components of quantitative HTA practice, even though the measurement principles required to justify them receive low endorsement. This is the central reversal. The outputs of the reference case are accepted, while the conditions necessary to validate those outputs are neglected.

The QALY results are particularly important. The statement that the QALY is a ratio measure receives probabilities from 0.85 to 0.90. The statement that the QALY is dimensionally homogeneous receives 0.85 to 0.90. QALYs can be aggregated receives 0.85 to 0.90. These findings indicate strong reinforcement of the standard reference-case treatment of the QALY. Yet the QALY depends on multiplying a utility value by time. Time is correctly recognized as a ratio measure, with a probability of 0.95, but the corresponding requirement that the utility component also possess ratio properties is not recognized. The result is dimensional failure. One valid ratio component cannot rescue a second component whose measurement status has not been established.

The same problem applies to reference-case simulation. The statement that reference-case simulations generate falsifiable claims receives high endorsement in the reference-case direction, with probabilities between 0.85 and 0.90. Yet simulation models do not create measurement. They manipulate quantities already assumed to be valid. If utilities are not ratio measures, if QALYs are not dimensionally homogeneous, and if latent attributes are not measured, then simulation merely extends the defect across time, populations and scenarios. More elaborate modelling cannot compensate for invalid measurement inputs.

The moderate endorsement of falsification and logits is revealing. Non-falsifiable claims should be rejected receives probabilities from 0.60 to 0.65, and the logit as the natural logarithm of the odds ratio receives 0.60 to 0.65. These results suggest some residual awareness of empirical testing and statistical form. However, this recognition does not extend to the measurement requirements that would make claims testable. Falsification requires claims that are based on measures. If the underlying quantity is a numerical construction rather than a measure, the claim cannot be meaningfully tested as a quantitative proposition. Thus, even where scientific language is present, the measurement foundation remains absent.

The narrow ranges across the seven centers are among the most important findings. There is little evidence of meaningful variation in the structure of endorsement. The same TRUE propositions receive low endorsement. The same FALSE propositions receive high endorsement. The same Rasch and representational measurement concepts are almost absent. The same QALY and simulation assumptions are reinforced. This consistency indicates that measurement inversion is not a local weakness, an institutional anomaly or a difference in methodological emphasis. It is a consensus pattern across the US academic HTA environment.

The implications are substantial. These seven research centers represent influential sources of HTA teaching, research, methods development and professional training. If their knowledge bases give little recognition to representational measurement, ratio measurement, unidimensionality, dimensional homogeneity, Rasch measurement and latent attribute possession, then the individuals trained within these environments are unlikely to acquire the conceptual tools required to challenge the reference-case paradigm. Curriculum inversion and measurement inversion therefore reinforce one another. Students and researchers are trained to apply utilities, QALYs, cost-effectiveness analysis and simulation models without first being equipped to determine whether the quantities involved are measures.

The findings also explain why the reference-case paradigm has persisted. Its survival should not be mistaken for scientific validation. The interrogation results suggest a different interpretation: the concepts necessary to expose the measurement failure are largely absent from the knowledge base. A framework built on arithmetic before measurement can persist when the educational and research environment does not teach the principles required to recognize the inversion. The persistence of the paradigm therefore reflects institutional reproduction rather than scientific legitimacy.

The conclusion is unavoidable. The seven US research-center interrogations demonstrate a stable and coherent pattern of measurement inversion. Foundational principles of measurement science are weakly endorsed, while the assumptions required to sustain utilities, QALYs and reference-

case simulations are strongly reinforced. The pattern is not transitional, ambiguous or institution-specific. It is a consensus profile. For a discipline claiming to produce robust estimates of therapeutic value and therapy impact, this is devastating. The results indicate that the contemporary US HTA academic knowledge base remains anchored in a paradigm whose central constructs fail the standards of measurement required for quantitative science.

FROM MEASUREMENT INVERSION TO CURRICULUM INVERSION

The results of the seven United States research-center interrogations have profound implications for the future of health technology assessment. The evidence demonstrates a consistent pattern of measurement inversion. Foundational principles of representational measurement receive little endorsement, while propositions embedded in the utility, QALY, and reference-case simulation framework receive strong endorsement. The findings are remarkably consistent across institutions. The issue is therefore not the methodological preferences of individual centers but the existence of a shared intellectual framework that has become detached from the standards governing quantitative science.

This immediately raises a more fundamental question. What is driving measurement inversion? How has a discipline come to rely so heavily upon utilities, QALYs, cost-effectiveness ratios, and simulation models while simultaneously giving so little recognition to measurement, scale theory, unidimensionality, dimensional homogeneity, and Rasch measurement? The most plausible explanation lies in the educational environment itself.

If researchers, analysts, and policy makers are not exposed to the concepts necessary to evaluate quantitative claims, then measurement inversion becomes unsurprising. A discipline cannot be expected to recognize deficiencies in measurement if the concepts required to identify those deficiencies are absent from its educational foundations. The issue therefore shifts from research outputs to the curriculum that produces those outputs. Measurement inversion invites a second interrogation: curriculum inversion.

Curriculum inversion occurs when students are taught how to perform quantitative analyses without first being taught the measurement principles that determine whether those analyses are scientifically legitimate. Economic evaluation, utility assessment, QALY construction, and simulation modelling become the focus of instruction, while the conceptual foundations upon which quantitative claims depend receive little attention. Students learn how to manipulate numbers before they learn whether those numbers represent measures. Arithmetic precedes measurement in education just as it precedes measurement in practice.

The implications are serious. A curriculum affected by inversion is not merely incomplete; it is potentially misleading. Students may emerge with considerable technical expertise in modelling and economic evaluation while lacking the conceptual framework necessary to evaluate the scientific legitimacy of the methods they have been taught. Concepts such as representational measurement, admissible arithmetic, ratio measurement, manifest and latent attributes, dimensional homogeneity, unidimensionality, and Rasch measurement become peripheral or disappear entirely. As a consequence, the assumptions underlying the reference-case paradigm are reproduced without critical examination.

Once attention is directed to these omissions, the long-term sustainability of the existing curriculum becomes questionable. A curriculum that neglects the principles governing measurement cannot indefinitely support claims to scientific rigor. The same standards that expose measurement inversion in research practice inevitably expose curriculum inversion in professional education.

The question then becomes how curriculum content should be assessed. A companion analysis applies the same large-language-model interrogation framework used in the measurement inversion studies but focuses on curriculum exposure statements rather than measurement propositions. Instead of asking whether a knowledge base endorses particular measurement assumptions, the curriculum interrogation asks whether students are likely to encounter the concepts necessary to evaluate quantitative claims. The objective is straightforward: to determine whether the educational environment equips future HTA practitioners with the intellectual tools required for measurement-based science or merely trains them in the application of the existing reference-case framework.

In this way, curriculum inversion provides the missing explanation for measurement inversion. The former reproduces the latter. Together they reveal how a framework built on arithmetic before measurement could become institutionalized and sustained across successive generations of researchers, educators, and decision makers.

NOT FIT FOR PURPOSE

The results presented in this assessment stand as a devastating critique of the contemporary HTA knowledge base across the seven United States research centers. The issue is not that particular methods require refinement, nor that individual assumptions might benefit from revision. The issue is that the intellectual foundations necessary to support quantitative claims regarding therapy impact are largely absent. Concepts central to representational measurement, ratio measurement, unidimensionality, admissible arithmetic, dimensional homogeneity, manifest and latent attributes, and Rasch measurement receive little recognition, while propositions embedded in the utility, QALY, and reference-case simulation framework are strongly reinforced. The conclusion is unavoidable: the present educational and research framework is not fit for purpose.

This finding has immediate implications for HTA education. If the objective of an HTA curriculum is to prepare students to evaluate the therapeutic impact and value of competing interventions, then the curriculum must first address the question of measurement. Before students can assess value, compare outcomes, construct economic evaluations, or make claims regarding comparative effectiveness, they must understand the conditions under which quantitative claims are possible. Without measurement, there can be no meaningful arithmetic. Without admissible arithmetic, there can be no scientifically credible claims regarding therapy impact.

The starting point for curriculum reconstruction is therefore straightforward. If quantitative assessment is to satisfy the standards of normal science, it must be based upon ratio measurement. Once this principle is accepted, the range of legitimate measurement options narrows dramatically. There are only two classes of ratio measure available for HTA. Manifest attributes, those that are directly observable, require linear ratio measures. Examples include survival time, hospital

admissions, emergency department visits, treatment persistence, medication possession, adverse events, and other empirically observable outcomes. These attributes can support lawful arithmetic because their measurement properties can be demonstrated directly.

Latent attributes present a different challenge. Pain, fatigue, depression, functioning, treatment satisfaction, quality of life, and similar constructs are not directly observable. They cannot be measured through simple observation because the attribute itself is hidden. The required outcome is not a score, an index, or a utility value, but possession of the latent attribute. In these circumstances the only scientifically defensible measurement framework is the Rasch model, which produces a Rasch logit ratio measure. This provides the necessary conditions for unidimensionality, invariance, and lawful measurement of latent trait possession.

The implications for HTA education are profound. Much of the existing curriculum is devoted to techniques that presuppose measurement rather than demonstrating it. Students are taught utility assessment, QALY construction, cost-effectiveness analysis, and simulation modelling without first establishing whether the quantities entering those calculations satisfy the requirements of measurement. The result is a curriculum focused on numerical manipulation rather than measurement science.

A curriculum designed to support scientific HTA would look very different. It would begin with attributes. Students would learn how to identify the target attribute, determine whether it is manifest or latent, establish the appropriate measurement framework, and construct evaluable and falsifiable claims. Arithmetic would follow measurement rather than precede it. Utilities, QALYs, and simulation models would no longer occupy the center of the curriculum because they would be recognized as numerical constructions lacking the measurement foundations required to support quantitative claims.

The findings from the seven US research-Centre interrogations therefore point to a stark conclusion. The current HTA curriculum does not merely contain gaps; it is structured around assumptions that fail the standards of measurement science. A curriculum built upon arithmetic before measurement cannot prepare students to undertake scientific assessment of therapeutic value. Reconstruction is therefore not optional. If HTA is to retain intellectual credibility, education must be rebuilt around the only two measurement frameworks capable of supporting quantitative claims: linear ratio measures for manifest attributes and Rasch logit ratio measures for latent attributes.

FROM PARADIGM FAILURE TO PARADIGM RECONSTRUCTION

After more than forty years, it will come as a surprise to many researchers, educators, and policy makers to learn that the contemporary HTA paradigm, with its reliance on utilities, QALYs, cost-effectiveness ratios, and reference-case simulation models, fails to satisfy the standards required by measurement theory and representational measurement. This conclusion is particularly troubling because the relevant standards were not newly discovered. The principles governing measurement scales, ratio measurement, admissible arithmetic, unidimensionality, and representational measurement were established decades before the emergence of the reference-case framework. The issue is therefore not that HTA failed to keep pace with developments in

measurement science. Rather, it appears that the discipline developed largely without recognizing the implications of those standards for its own analytical practices.

The findings presented here are therefore devastating. They demonstrate that a subject area long regarded as scientifically rigorous and methodologically sophisticated rests upon foundations that cannot support the quantitative claims made on its behalf. Utilities cannot be assumed to possess ratio properties. QALYs cannot be assumed to satisfy dimensional homogeneity. Simulation models cannot create measurement where measurement does not exist. The consequence is that many of the central constructs of contemporary HTA are revealed as numerical constructions rather than measures. What appeared to be a settled scientific framework is now exposed as a paradigm whose measurement foundations were never secured.

How such a situation persisted for four decades remains an important question. The most plausible explanation is curriculum inversion. If students, researchers, and policy analysts are not taught representational measurement, scale theory, ratio measurement, manifest and latent attributes, and Rasch measurement, then they are unlikely to recognize the deficiencies of the framework they inherit. The present curriculum becomes self-replicating. Each generation learns the methods of the reference case while remaining largely unaware of the measurement principles necessary to evaluate those methods. Whether curriculum inversion is the sole explanation remains open to debate, but the evidence strongly suggests that it has played a central role in sustaining measurement inversion across research, education, and policy.

The implications are clear. The existing paradigm cannot be preserved through incremental reform. Once measurement inversion has been identified, there is no credible path back to the status quo. The challenge is not to repair the reference-case framework but to replace it with one that satisfies the standards of normal science. If HTA is to remain relevant as a discipline concerned with therapeutic value and therapy impact, it must be rebuilt upon lawful measurement rather than numerical construction.

Fortunately, the path forward is already available. The question is no longer what a scientifically credible HTA curriculum should contain. The essential elements have been identified and assembled within the nine-unit Maimon Research LLC reconstruction program. This provides a comprehensive educational framework for replacing the traditional reference-case paradigm with one grounded in representational measurement and the standards of normal science. Beginning with the principle that measurement must precede arithmetic, the program introduces attributes, scale theory, ratio measurement, manifest and latent attributes, and the role of Rasch logit ratio measurement for latent outcomes. It demonstrates why utilities, QALYs, and reference-case simulation models fail the requirements for lawful quantitative claims and presents an alternative framework based on evaluable and falsifiable assessments of therapeutic impact. Designed for students, faculty, researchers, manufacturers, and HTA agencies, the program serves as a practical resource for curriculum reconstruction and the transition to measurement-based health technology assessment. Details are given on the Maimon Research website ¹¹ .

CONCLUSIONS

The findings reported in this assessment leave little room for ambiguity. Across seven leading United States HTA research centers, the interrogation results reveal a stable and remarkably consistent pattern of measurement inversion. The foundational principles required for quantitative science receive weak endorsement, while the assumptions that sustain utilities, QALYs, cost-effectiveness ratios, and reference-case simulation models receive strong reinforcement. This pattern is not confined to a single institution, methodological tradition, or research group. It is a consensus profile that characterizes the contemporary US HTA knowledge base.

The significance of these findings extends far beyond methodological criticism. The issue is not whether existing HTA methods require refinement or whether additional statistical sophistication might improve current practice. The issue is that the measurement foundations necessary to support quantitative claims have never been secured. Concepts such as representational measurement, ratio measurement, admissible arithmetic, dimensional homogeneity, unidimensionality, manifest and latent attributes, and Rasch measurement remain largely absent from the intellectual framework within which HTA operates. The consequence is that arithmetic has been allowed to assume the role that properly belongs to measurement. Numerical constructions are treated as measures, while the requirements that distinguish measures from numbers remain largely unexamined.

The evidence also points to a deeper problem. Measurement inversion does not emerge spontaneously. It is reproduced through curriculum inversion. Students, researchers, analysts, and policy makers are trained within educational environments that emphasize utility assessment, QALY construction, cost-effectiveness analysis, and simulation modelling while giving little attention to the principles required to determine whether these activities are scientifically legitimate. The result is a self-sustaining system in which the assumptions of the reference-case paradigm are transmitted from one generation to the next without exposure to the conceptual tools needed to challenge them. The persistence of the paradigm therefore reflects institutional reproduction rather than scientific validation.

The implications for HTA are profound. Once measurement inversion and curriculum inversion have been identified, there can be no return to the status quo. The contemporary framework cannot be rescued through revised utility instruments, updated methodological guidance, more sophisticated simulation models, or increasingly elaborate statistical techniques. None of these responses address the underlying problem. Arithmetic cannot create measurement where measurement does not exist. The challenge is therefore not one of reform but of reconstruction.

The choice facing HTA is therefore stark. One option is to continue defending a paradigm whose central constructs fail the standards required for quantitative science. The other is to acknowledge the implications of measurement inversion and curriculum inversion and undertake the transition to a framework grounded in lawful measurement. The findings reported here suggest that the future credibility of HTA will depend upon that choice. If reconstruction is embraced, HTA may yet emerge as a discipline consistent with the standards of normal science.

If it is resisted, the relevance and legitimacy of the reference-case paradigm will continue to erode as awareness of its measurement deficiencies becomes increasingly widespread.

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 United States: The absence of representational measurement and the Schaeffer Center for Health Policy and Economics, University of Southern California. Logit Working Paper 750. <https://maimonresearch.com/logit-working-paper-no-750-april-2026/>
- ² United States: The absence of representational measurement and the Center for Health Decision Science, Harvard University. Logit Working Paper No 751. <https://maimonresearch.com/logit-working-paper-no-751-april-2026/>
- ³ United States: The absence of representational measurement and the University of Washington CHOICE Institute. Logit Working Paper No 752. <https://maimonresearch.com/logit-working-paper-no-752-april-2026/>
- ⁴ United States: The absence of representational measurement and the Duke-Margolis Institute for Health Policy, Duke University. Logit Working Paper No 753. <https://maimonresearch.com/logit-working-paper-no-753-april-2026/>
- ⁵ United States: The absence of representational measurement and the Leonard Davis Institute of Health Economics, University of Pennsylvania. Logit Working Paper No 754. <https://maimonresearch.com/logit-working-paper-no-754-april-2026/>
- ⁶ United States: The absence of representational measurement and the Health Economics Group, Bloomberg School of Public Health. Johns Hopkins University. Logit Working Paper No 756. <https://maimonresearch.com/logit-working-paper-no-756-april-2026/>
- ⁷ United States: The absence of representational measurement and the Stanford University Health Policy Center. Logit working Paper No 755. <https://maimonresearch.com/logit-working-paper-no-755-april-2026/>
- ⁸ Stevens S. On the Theory of Scales of Measurement. *Science*. 1946;103(2684):677-80
- ⁹ Krantz D, Luce R, Suppes P, Tversky A. Foundations of Measurement Vol 1: Additive and Polynomial Representations. New York: Academic Press, 1971

¹⁰ Bond T, Zi Yan, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences (4th Ed). New York Routledge, 2021

¹¹ Maimon Research Transition Program <https://maimonresearch.com/hta-reconstruction-program-and-fees/>