ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: MEASUREMENT INVERSION AND THE NICE SCIENCE POLICY AND RESEARCH PROGRAM

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

LOGIT WORKING PAPER No 687 JUNE 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

For more than two decades, the National Institute for Health and Care Excellence (NICE) has occupied a leading position in the development of health technology assessment (HTA), with its reference-case framework exerting substantial influence on reimbursement policy, academic research, and HTA methodology internationally. Central to this framework are utilities, quality-adjusted life years (QALYs), cost-effectiveness ratios, and lifetime simulation models. The present study asks a fundamental scientific question: does the NICE Science Policy and Research (SP&R) Programme, the methodological research arm responsible for developing and refining NICE methods, recognize the principles of representational measurement that must precede all quantitative claims?

Using a large language model interrogation of the SP&R Program knowledge base, twenty-four canonical statements derived from representational measurement theory, scale theory, Rasch measurement, and the philosophy of scientific inference were evaluated through categorical endorsement probabilities and normalized logit scores. The interrogation examined whether the methodological foundations of the NICE reference case recognize the requirements of unidimensionality, ratio measurement, dimensional homogeneity, admissible arithmetic, manifest and latent attributes, and evaluable and falsifiable therapy impact claims.

The results demonstrate a consistent pattern of measurement inversion. Statements representing the scientific conditions required for quantitative measurement receive uniformly weak endorsement, while propositions supporting utilities, QALYs, and reference-case simulation models receive consistently strong endorsement. The interrogation reveals a series of internal contradictions in which the scientific conditions necessary to justify arithmetic operations are largely rejected while the numerical constructions dependent upon those conditions are strongly endorsed. Particular attention is given to the EQ-5D valuation algorithm, the construction of utilities, QALYs, cost-effectiveness ratios, and reference-case simulation models, demonstrating that each stage inherits unresolved failures of representational measurement. The analysis identifies successive failures of measurement, arithmetic, and inference that cannot be remedied by increasingly sophisticated econometric modelling or simulation techniques.

The conclusion is uncompromising. The NICE reference-case framework did not evolve into scientific inadequacy; it failed to satisfy the required standards of quantitative measurement from its inception. Consequently, the challenge facing NICE is not further methodological refinement but methodological reconstruction. Scientific HTA must be rebuilt upon representational measurement, lawful ratio measurement for both manifest and latent attributes, admissible arithmetic, and prospectively specified, evaluable, replicable, and falsifiable claims regarding therapy impact. Only under these conditions can HTA claim membership within the quantitative sciences.

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 Kingdom 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 more than 240 interrogations of global 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.

The purpose of the present study is to extend a recent assessment of the observed pattern of measurement and curriculum inversion from a knowledge base of five leading United Kingdom HTA research centers: the Centre for Health Economics (CHE), the Centre for Reviews and Dissemination (CRD), the Oxford Health Economics Research Centre and HTA Group, the School of Health and Related Research (ScHARR), and the Newcastle Institute of Health and Society to the national Institute of Health and Care Excellence; specifically the NICE Science Policy and Research P:rogram (SP&R) [1 2] .

The SP & R program provides the methodological research that underpins the continued development of NICE guidance and health technology assessment. Its role is to identify methodological challenges arising from new technologies, evidence sources, and decision-making requirements, and to commission or undertake research that strengthens NICE evaluation methods. The program supports the refinement of evidence synthesis, economic evaluation, modelling approaches, real-world evidence, and health-related quality-of-life assessment, while identifying priorities for future methodological research. As the principal source of methodological development within NICE, the SP&R Program plays a central role in shaping the analytical

framework that underpins technology appraisals, clinical guidelines, and health technology evaluations across the organization.

The key question, therefore, is whether the SP&R Program endorses measurement inversion. If it does, this would be entirely consistent with the findings for the leading UK HTA research centers and with every one of the 240 large language model interrogations undertaken to date. As the methodological research arm of NICE, the SP&R Program occupies a pivotal position in the development and refinement of the methods that underpin the NICE reference-case framework. Consequently, its recognition or otherwise of the principles of representational measurement has implications extending far beyond the program itself.

If the SP&R Program knowledge base demonstrates the same limited recognition of the standards governing quantitative measurement as the leading UK research centers, then the conclusion is unavoidable. The problem is not confined to individual institutions but lies within the scientific foundations of the UK reference-case paradigm itself. The purpose of this assessment is therefore not simply to evaluate a single methodological program, but to determine whether the intellectual foundations that support NICE guidance remain consistent with the established requirements of quantitative science [3].

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 scales [4]. 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 measurement [5]. 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 scale [6]. 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.

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 knowledge base interrogated for this assessment comprised the publicly accessible methodological resources of the NICE Science Policy and Research (SP&R) Program. Unlike the broader NICE knowledge base, which encompasses health technology evaluations, clinical guidelines, evidence reviews, diagnostics, and implementation guidance, the SP&R Program is specifically responsible for the development, evaluation, and refinement of the methods that underpin NICE decision making. The interrogation therefore focused on documents describing methodological research priorities, evidence synthesis, economic evaluation, modelling, health-related quality-of-life measurement, real-world evidence, uncertainty analysis, and future methodological development. These materials were considered together with associated descriptions of the NICE reference-case framework and methodological guidance where they informed the program's research agenda.

The SP&R Program represents the methodological foundation upon which NICE continually develops and updates its assessment methods. Consequently, it provides an appropriate and representative knowledge base for evaluating whether the scientific principles governing quantitative measurement are recognized within the methodological framework itself. Of particular interest was the extent to which the program acknowledges representational measurement, the principal scales of measurement, ratio measurement, dimensional homogeneity, manifest and latent attributes, Rasch measurement, and the scientific requirements for evaluable and falsifiable claims regarding therapy impact. Because the program influences both NICE methodology and, indirectly, HTA research and education internationally, its knowledge base provides an important indicator of the scientific foundations of the contemporary NICE reference-case paradigm.

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 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 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
25.

REVIEW: NICE SP&R PROGRAM AND MEASUREMENT INVERSION

It is important to recognize that the SP&R Program exists to improve the methods that NICE uses to develop guidance and health technology assessments. For that reason, the program functions as a methodological gatekeeper. It does not merely reflect the prevailing HTA paradigm; it develops, refines, and legitimizes the methodological framework that underpins NICE guidance. Because NICE has long been regarded as one of the world's leading HTA agencies, its methodological influence extends well beyond the United Kingdom. NICE methods have shaped academic research, informed HTA guidance internationally, and provided a benchmark against which many national agencies have developed their own assessment frameworks. Consequently, if the SP&R Program fails to recognize the requirements of representational measurement, the implications extend beyond NICE itself (Table 1). They raise fundamental questions regarding the scientific foundations of one of the most influential methodological paradigms in contemporary health technology assessment.

TABLE 1: ITEM STATEMENT: RESPONSE AND ENDORSEMENT

STATEMENT	**RESPONSE 1=TRUE 0=FALSE**	**ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY**	**NORMALIZED LOGIT +/- 2.50**
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.70
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1,70
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.85	+1,70
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	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.75	+1.10
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.05	-2.50
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	0.75	+1.10
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.90	+2.20
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.70
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.60	+0.40
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	0.35	-0.60
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.10	-2.20

THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The NICE SP&R Program occupies a unique position within contemporary health technology assessment. Unlike an academic research center, whose principal responsibilities are research and teaching, or a journal, whose role is to disseminate scientific findings, the SP&R Program exists to develop, refine, and support the methods that NICE employs in producing health technology evaluations, clinical guidelines, and policy recommendations. Its influence therefore extends beyond methodological research itself. It functions as the principal methodological gatekeeper for NICE, determining not only which analytical approaches are considered acceptable but also identifying priorities for future methodological development.

Because NICE has, for more than two decades, been regarded internationally as one of the leading agencies in health technology assessment, the influence of the SP&R Program extends well beyond the United Kingdom. Its methodological assumptions are reflected in academic research, postgraduate education, journal publications, manufacturer submissions, and the development of HTA agencies throughout Europe, North America, Australasia, and many middle-income countries. An assessment of the SP&R Program knowledge base is therefore far more than an examination of a single methodological program. It is an assessment of the intellectual foundations supporting one of the world's most influential HTA paradigms.

The central question posed by the present interrogation is straightforward. Does the SP&R Program recognize the scientific standards governing quantitative measurement, or does it continue to endorse the reference-case framework that has characterized HTA for more than four decades? The results indicate a strikingly consistent pattern. Statements reflecting the principles of representational measurement receive uniformly low endorsement probabilities, while statements supporting utilities, quality-adjusted life years (QALYs), and reference-case simulation models receive consistently high endorsement. The resulting profile is not one of isolated omissions or occasional inconsistencies. It demonstrates a coherent pattern of measurement inversion in which arithmetic is systematically permitted to precede measurement.

Perhaps the single most revealing result concerns the statement which asserts that measurement must precede arithmetic. This proposition lies at the heart of every quantitative scientific discipline. Whether one is measuring length, mass, temperature, electrical resistance, or latent psychological attributes, arithmetic operations are performed only after the quantities concerned have been shown to satisfy the appropriate measurement conditions. The SP&R interrogation assigns this proposition an endorsement probability of only 0.10 (logit –2.20). Such a result indicates almost complete absence of recognition. This finding is particularly important because every subsequent quantitative operation within the NICE reference case depends upon this principle. Cost-utility analysis requires discounting of time by utility values. Cost-effectiveness ratios require division of monetary costs by QALYs. Incremental analyses require subtraction, aggregation, averaging, and comparison of these quantities. If measurement does not precede

arithmetic, then every subsequent analytical operation rests upon an unverified assumption. The interrogation therefore identifies the defining feature of measurement inversion at its very origin.

The same pattern is reinforced by the statement that arithmetic operations require satisfaction of the axioms of representational measurement. Again, the endorsement probability is only 0.05 (logit –2.50), representing virtually complete rejection. This result is remarkable because representational measurement has, for more than half a century, provided the accepted scientific framework governing quantitative measurement across the physical and social sciences. The axioms formalized by Krantz, Luce, Suppes, and Tversky were not proposed as optional methodological preferences but as the conditions necessary for assigning numbers to empirical phenomena in ways that preserve their relational structure. The absence of endorsement therefore suggests that the methodological research program responsible for improving NICE methods gives almost no recognition to the scientific framework that determines whether quantitative claims are lawful. Instead, numerical manipulation proceeds independently of the measurement conditions that should govern it.

The implications become even more apparent when the interrogation considers ratio measurement: the assertion that multiplication requires ratio measures. This proposition receives an endorsement probability of only 0.10 (logit –2.20). Yet multiplication lies at the very center of the NICE reference-case framework. The QALY is explicitly constructed by multiplying a measure of time by a health-state utility value. Time itself presents no difficulty. Time is recognized as a ratio measure, receiving almost complete endorsement with a probability of 0.95 (logit +2.50). Thus, the knowledge base recognizes the ratio properties of one component of the QALY while simultaneously failing to recognize that the second component must possess the same measurement properties before multiplication becomes scientifically admissible. This asymmetry is highly revealing. It demonstrates that the arithmetic operation is accepted without first establishing that both operands satisfy the dimensional homogeneity requirements imposed by representational measurement.

The contradiction becomes explicit when one considers the proposition that the QALY is a ratio measure and receives an endorsement probability of 0.90 (logit +2.20). The statement that the QALY satisfies dimensional homogeneity, receives an equally strong endorsement. These results stand in direct opposition to the near rejection of statements that multiplication requires a ratio measure, that measurement precedes arithmetic and that meeting the axioms of representational measurement is required for arithmetic. In effect, the interrogation reveals that the SP&R Program knowledge base strongly endorses the conclusions required for the operation of the reference case while simultaneously rejecting the scientific conditions necessary to justify those conclusions. This is not simply a matter of differing methodological opinion. It represents an internal inconsistency within the knowledge base itself. A framework cannot simultaneously deny the necessity of ratio measurement while affirming the legitimacy of multiplying time by utility values. The interrogation therefore identifies a contradiction at the center of the NICE methodological framework.

The same inconsistency is evident in relation to health-state utilities. The statement is endorsed that the EQ-5D preference algorithms create interval measures. Nevertheless, the interrogation indicates strong endorsement of the use of EQ-5D utilities within cost-utility analysis. The

distinction is important. The present critique does not challenge the usefulness of collecting patient or public preferences. Preference elicitation undoubtedly provides valuable descriptive information concerning perceptions of health states. The scientific question, however, is whether the resulting numerical values constitute lawful quantitative measures. The interrogation indicates that the SP&R Program knowledge base gives almost no recognition to the measurement principles required to answer that question while continuing to endorse the use of those values in arithmetic operations fundamental to the reference case.

This result illustrates the broader phenomenon of numerical substitution. Numbers are treated as though they automatically possess measurement properties merely because they are numerical. Once preference values have been generated, they become candidates for multiplication, averaging, subtraction, aggregation, and incorporation into increasingly sophisticated simulation models. Yet representational measurement makes no such assumption. Numbers acquire scientific legitimacy only when they preserve the empirical relationships characterizing the attribute under investigation. Without that prior demonstration, arithmetic becomes an exercise in numerical manipulation rather than quantitative science. The SP&R Program interrogation indicates that this distinction is almost entirely absent from the NICE methodological knowledge base.

The interrogation also demonstrates remarkably limited recognition of latent measurement. Statements addressing the role of Rasch measurement in transforming ordinal observations into lawful measures of latent attributes, receive endorsement probabilities of only 0.05 (logit –2.50). These are among the lowest endorsement scores observed in the entire interrogation. This finding has profound implications because many outcomes central to health technology assessment, pain, fatigue, functional status, depression, social participation, and health-related quality of life, are latent rather than manifest attributes. They cannot be observed directly in the manner of body weight, survival time, or hospital admissions. Their measurement therefore requires a formal measurement model capable of establishing a quantitative continuum representing possession of the latent attribute. The interrogation suggests that the NICE methodological framework gives virtually no recognition to this requirement, despite its central importance for patient-reported outcomes and quality-of-life assessment.

The absence of Rasch measurement is not a minor methodological omission. It reveals the broader philosophical orientation of the SP&R Program knowledge base. Rather than asking whether latent attributes have first been measured, the NICE framework proceeds directly to the valuation of health states and the construction of utility scores. In effect, preference elicitation replaces measurement. The consequence is that ordinal observations and social preferences become the foundation for arithmetic operations without any prior demonstration that the underlying attribute satisfies the requirements of representational measurement. This explains why the interrogation assigns almost complete rejection to classes of measurement, Rasch rules and interval measurement and summation of Likert scores as ratio measures while simultaneously giving strong endorsement to the utility-based framework that dominates NICE guidance. Once again, the interrogation exposes a systematic inversion of scientific reasoning. Measurement is assumed rather than demonstrated.

The interrogation also highlights an equally important omission concerning attribute structure. The statement asserting that measures must be unidimensional, receives an endorsement probability

of only 0.15 (logit –1.70). Yet unidimensionality represents one of the fundamental conditions for quantitative measurement. Unless an instrument measures a single attribute, no numerical assignment can legitimately claim to represent variation in that attribute. Contemporary utility instruments combine multiple domains of mobility, self-care, pain, anxiety, usual activities and other components into composite preference scores. Such instruments may provide useful descriptive summaries, but the interrogation suggests that the SP&R Program knowledge base gives little recognition to the scientific distinction between multidimensional descriptive profiles and quantitative measures of a single attribute. The resulting composite scores are then treated as though they possess the properties required for arithmetic. This again reflects measurement inversion rather than representational measurement.

The interrogation therefore identifies a recurring pattern of internal contradiction. Statements expressing the scientific conditions necessary for lawful quantitative claims receive consistently low endorsement, whereas propositions required to sustain the NICE reference case receive consistently high endorsement. The endorsement rejects the necessity for ratio measurement before multiplication, yet there is the strongly endorsed QALY as though it were a ratio measure. There is no recognition of the axioms of representational measurement, while strongly endorsing the dimensional legitimacy of the QALY. Rasch measurement is rejected as the basis for latent measurement, yet patient-reported outcomes and health-related quality of life continue to occupy central positions within NICE assessments. These are not isolated anomalies. Together they demonstrate that the reference-case paradigm depends upon propositions that cannot be reconciled with the scientific principles governing quantitative measurement.

A key question is not whether modelling can be improved. The NICE SP&R Program devotes considerable effort to refining evidence synthesis, econometric methods, simulation techniques, structural assumptions, uncertainty analysis, and model validation. These developments may improve the internal consistency or computational performance of the reference-case framework. They do not, however, address the prior scientific question: do the quantities entering the model satisfy the established standards of representational measurement? Unless that question is answered in the affirmative, improvements in modelling cannot confer scientific legitimacy upon the model's outputs. Mathematical sophistication cannot compensate for failure of measurement because every model inherits the measurement properties of its inputs. If the inputs are not lawful measures, the outputs cannot become lawful measures merely because increasingly sophisticated analytical techniques have been applied. The reference case has concentrated on improving arithmetic when the unresolved problem is measurement.

Scientific models do not create measurement. They operate upon measurements that have already been established. Newton's equations do not determine the measurement of mass or distance. Einstein's field equations did not create the concepts of length, time or mass-energy. They relied upon measurement systems whose properties had already been established independently. Mathematical sophistication cannot compensate for defective measurement because every model inherits the properties of the quantities supplied to it. If the inputs fail to satisfy representational measurement, no degree of computational sophistication can transform them into lawful measures. The interrogation therefore identifies the central weakness of the NICE methodological framework. Considerable intellectual effort has been devoted to improving models while

comparatively little attention has been given to the scientific legitimacy of the quantities entering those models.

The issue is therefore not assumptions themselves. Every scientific model contains assumptions. Rather, the issue concerns the status of the quantities to which those assumptions are applied. NICE devotes considerable attention to sensitivity analyses, probabilistic analyses, validation exercises and structural uncertainty. These procedures examine how model outputs respond to changes in assumptions. They do not establish that the quantities entering the model possess lawful measurement properties. Consequently, they cannot resolve measurement inversion. They simply explore alternative consequences of assumptions imposed upon quantities whose measurement status remains unresolved.

This distinction is reflected in the interrogation results relating to falsifiability. The statement that non-falsifiable claims should be rejected, receives only moderate endorsement while the statement rejecting the proposition that reference-case simulations generate falsifiable claims, receives strong endorsement of the contrary position. This reveals another important characteristic of the NICE methodological framework. Simulation models generate conditional projections rather than prospectively specified empirical claims capable of direct refutation. Their outputs depend upon assumptions concerning disease progression, treatment persistence, transition probabilities, utility values, costs, discount rates and numerous other modelling choices. If observations subsequently differ from predictions, the explanation is almost invariably sought through modification of assumptions rather than rejection of the underlying framework. Consequently, simulation modelling lacks the decisive empirical confrontation that characterizes scientific hypothesis testing in the Popperian tradition [7] [8]

This observation is particularly important because the SP&R Program is explicitly concerned with improving methods. Scientific progress, however, does not occur simply by refining analytical techniques. It also requires willingness to reconsider the assumptions upon which those techniques rest. The interrogation suggests that the SP&R Program knowledge base focuses overwhelmingly upon methodological refinement within the existing reference-case paradigm rather than questioning whether the paradigm itself satisfies the conditions required for quantitative science. In this respect the program functions less as an evaluator of methodological foundations than as their custodian. It improves the machinery without examining whether the machinery is built upon scientifically defensible measurements.

The institutional implications are considerable. NICE occupies a position of exceptional authority within international HTA. Manufacturers construct evidence submissions around NICE requirements. Academic research centers design methodological research with NICE guidance in mind. Journals publish studies employing NICE-compatible methods. Universities teach cost-utility analysis, QALYs and reference-case modelling because these methods are required in practice. Consequently, the SP&R Program does not merely influence NICE guidance. It influences the entire intellectual ecosystem surrounding health technology assessment. When the interrogation identifies measurement inversion within the SP&R Program knowledge base, it is identifying a mechanism through which measurement inversion is reproduced across research, teaching, publication and policy development.

This observation also explains the relationship between measurement inversion and curriculum inversion. The SP&R Program is not itself a teaching program. Nevertheless, it exerts profound influence upon university curricula because the methodological framework it develops becomes the framework that research centers teach. Students entering graduate programs in health economics and HTA are introduced to utilities, QALYs, economic evaluation and simulation modelling because these methods dominate NICE guidance and contemporary practice. Rarely are they first introduced to representational measurement, admissible arithmetic, dimensional homogeneity, the distinction between manifest and latent attributes, or Rasch measurement. Curriculum inversion therefore follows naturally from measurement inversion. Educational programs reproduce the assumptions already institutionalized within the methodological framework.

The interrogation consequently has implications extending well beyond NICE itself. It suggests that the persistence of the reference-case paradigm cannot be explained simply by methodological preference or historical accident. Rather, it reflects a process of institutional reproduction in which methodological assumptions become embedded within guidance, research priorities, educational programs, journal publications and professional expectations. Each component reinforces the others. Researchers employ the methods expected by NICE. Universities teach the methods employed by researchers. Journals publish studies conforming to prevailing methodological standards. Manufacturers prepare submissions consistent with NICE requirements. The resulting system possesses considerable internal stability despite its limited recognition of the scientific principles governing quantitative measurement.

It is therefore important to emphasize that the present critique does not question the commitment of the SP&R Program to methodological excellence within the existing paradigm. On the contrary, the program demonstrates substantial sophistication in evidence synthesis, decision modelling, statistical analysis and policy evaluation. The interrogation addresses a different question. Does the program recognize the scientific conditions that must be satisfied before arithmetic operations become lawful? The results indicate that it does not. Methodological sophistication and scientific measurement are not synonymous. Increasingly elaborate analytical techniques cannot compensate for failure to establish lawful measures. This distinction represents the central conclusion arising from the interrogation.

Taken as a whole, the endorsement profile demonstrates a coherent and internally consistent pattern. Statements reflecting the principles of representational measurement, ratio measurement, unidimensionality and Rasch measurement receive consistently low endorsement. Statements supporting utilities, QALYs, aggregation and reference-case modelling receive consistently high endorsement. The resulting profile is precisely that expected of a knowledge base committed to the reference-case paradigm rather than to representational measurement.

The NICE SP&R Program therefore does not merely reflect measurement inversion; it institutionalizes it. Because the program occupies one of the most influential methodological positions in international health technology assessment, this institutionalization extends far beyond the United Kingdom. It continues to shape research agendas, university curricula, evidence submissions and policy development throughout the global HTA community. The conclusion is therefore unavoidable. The challenge facing NICE is not further refinement of the reference case

but reconstruction of its methodological foundations around the established principles of representational measurement, lawful ratio measurement, explicit distinction between manifest and latent attributes, Rasch logit ratio measurement for latent outcomes, and prospectively specified, evaluable, replicable and falsifiable claims regarding therapy impact. Only by restoring measurement to its proper precedence over arithmetic can the methodological framework underpinning NICE guidance satisfy the standards expected of a quantitative scientific discipline.

IMPLICATIONS FOR THE EQ-5D-3L/5L

The extent of measurement inversion has profound implications for the two multiattribute instruments that underpin contemporary health technology assessment, the EQ-5D-3L and EQ-5D-5L. Once the principles of representational measurement are ignored, these instruments lose any claim to quantitative measurement. Their outputs are numerical constructions rather than lawful quantitative measures. Two fundamental errors are responsible for this conclusion: the absence of unidimensionality and the absence of lawful ratio measurement.

The first failure is foundational. Representational measurement requires that every quantitative claim refer to a single, unidimensional attribute. The EQ-5D does not measure a single attribute. It combines multiple dimensions of health, including mobility, self-care, usual activities, pain or discomfort, and anxiety or depression into composite health-state descriptions. Valuation of these multidimensional descriptions produces no more than a composite ordinal preference score. Such scores provide an ordering of preferences but possess none of the properties required for quantitative measurement. They are not measures of a single attribute, they do not possess a true zero, and they cannot support arithmetic operations beyond simple ordering. This initial failure is decisive because every subsequent analytical operation inherits the absence of lawful measurement.

The critical step in the construction of the EQ-5D utility is the valuation algorithm itself. Responses to the five health-status dimensions are first represented as a descriptive health-state profile. An econometric model is then used to estimate a series of coefficients—commonly referred to as utility weights—from population valuation data. These estimated regression coefficients are subsequently applied through a predefined valuation algorithm to generate a single composite utility score, typically expressed as a decrement from unity representing notional perfect health. The resulting value is then treated as though it were a lawful ratio measure of health-related quality of life.

The difficulty is that neither the valuation algorithm nor the resulting utility score has any recognized defense in terms of the axioms of representational measurement. The estimated coefficients are regression parameters, not measurement weights. They summarize statistical relationships within the valuation data but do not establish that the health-state responses constitute a unidimensional attribute, that the algorithm preserves the empirical relational structure of that attribute, or that the resulting utility possesses ratio-scale properties. The valuation algorithm is therefore a numerical construction rather than a measurement procedure.

Each utility value is consequently no more than a composite score derived from the arithmetic combination of ordinal health-state responses using estimated regression coefficients. The

valuation algorithm cannot transform ordinal observations into a ratio measure because arithmetic cannot create measurement. Measurement must already exist before arithmetic becomes scientifically admissible. There is no demonstration that the algorithm satisfies the axioms of representational measurement, establishes unidimensionality, preserves the empirical relational structure of the underlying attribute, produces dimensional homogeneity, or generates quantities possessing a true zero that support multiplication and division. Consequently, every utility generated by the algorithm remains a composite numerical construct with no established measurement properties.

The second failure concerns the application of these utility values in the construction of QALYs. Even if one were to accept the utility values generated by the valuation algorithm, multiplication is scientifically admissible only when both operands are ratio measures and when the resulting product satisfies dimensional homogeneity. Survival time satisfies these conditions. Utility values do not. Consequently, the arithmetic operation required to generate the QALY is itself inadmissible. The QALY is therefore not a lawful quantitative measure but the numerical consequence of multiplying a ratio measure by a composite ordinal score whose measurement properties have never been established.

These failures propagate throughout the remainder of the reference-case framework. Cost-effectiveness ratios divide monetary costs by QALYs whose measurement properties have not been demonstrated. Lifetime simulation models repeatedly aggregate, average, discount, extrapolate, and compare these quantities through increasingly sophisticated mathematical procedures. Yet every stage inherits the measurement deficiencies present at the beginning of the analytical sequence. Neither econometric modelling, valuation algorithms, simulation modelling, nor decision thresholds can compensate for the absence of representational measurement. The reference-case framework therefore exhibits three successive failures. The first is measurement failure, where lawful quantitative measurement is never established. The second is arithmetic failure, where inadmissible mathematical operations are performed on quantities lacking the required scale properties. The third is inference failure, where the resulting numerical constructions are interpreted as scientific evidence for reimbursement and policy decisions despite the absence of lawful measurement. Every subsequent stage inherits the deficiencies of those preceding it.

The endorsements of the false propositions detailed in Table 1 reinforce rather than weaken these conclusions. They are not isolated errors but components of a coherent methodological framework in which the scientific conditions governing quantitative measurement receive little recognition while the numerical constructions that depend upon those conditions receive strong endorsement. Thus, the proposition that the QALY is a ratio measure receives an endorsement probability of 0.90 (logit +2.20), despite the almost complete absence of endorsement for the principle that multiplication requires ratio measurement ($p = 0.10$, logit –2.20). Similarly, the proposition that the QALY satisfies dimensional homogeneity receives equally strong endorsement ($p = 0.90$, logit +2.20), while the proposition that arithmetic must be preceded by representational measurement receives virtually no recognition ($p = 0.05$, logit –2.50). The same contradiction is evident in the strong endorsement of utility-based arithmetic despite the failure to recognize that health-state valuations produce no more than composite ordinal scores whose measurement properties have never been demonstrated. These endorsement profiles reveal an internally inconsistent knowledge

base that simultaneously rejects the scientific conditions required for quantitative measurement while strongly endorsing the numerical constructions that depend upon those very conditions.

Taken together, these findings demonstrate far more than methodological disagreement. They show that the NICE reference-case framework rests upon a sequence of arithmetic operations that cannot be justified by the established principles of representational measurement. Every major stage in the analytical sequence from multidimensional health-state descriptions, through econometric estimation, utility construction, QALY generation, cost-effectiveness ratios, and lifetime simulation modelling depends upon quantities whose measurement properties have never been demonstrated. The issue is therefore not whether the reference-case framework can be improved through better econometric models, revised valuation algorithms, or more sophisticated simulation techniques. The issue is whether a framework built upon successive violations of lawful measurement can continue to claim scientific legitimacy. The present interrogation suggests that it cannot. Scientific HTA must begin where every quantitative science begins: with lawful measurement, followed by admissible arithmetic, and finally by prospectively specified, evaluable, replicable, and f

Finally, it might be argued that this critique misunderstands the purpose of the reference-case framework. The objective, it could be claimed, is not to generate scientifically testable hypotheses regarding therapy impact but simply to provide decision makers with a structured framework for making reimbursement decisions under conditions of uncertainty. Even if this more limited objective is accepted, the central criticism remains unchanged. Decision-support models are still quantitative models, and the quantitative operations they perform remain subject to the same principles of representational measurement that govern every other quantitative discipline. A simulation model cannot be exempted from the standards governing admissible arithmetic simply because its purpose is to inform policy rather than test scientific theory.

Indeed, if the reference case is intended as a rational aid to decision making, the requirement for lawful measurement becomes even more important. Decisions affecting the allocation of scarce healthcare resources should not rest upon arithmetic performed on quantities whose measurement properties have never been established. The issue is therefore not whether reference-case simulation has potential value as a decision-support framework. It may well have such value if constructed upon lawful measures. The difficulty is that the contemporary reference case begins with utilities that are no more than composite ordinal scores, constructs QALYs through inadmissible multiplication, derives cost-effectiveness ratios through inadmissible division, and then propagates these quantities through increasingly sophisticated simulation models. The resulting outputs cannot acquire scientific or quantitative legitimacy merely because they are intended to support policy decisions. A decision framework built upon quantities that fail the established standards of measurement cannot be rescued by appealing to its intended purpose. The prerequisite for both scientific inference and rational decision making is the same: measurement must precede arithmetic.

CONCLUSION: NICE AND PARADIGM FAILURE

The interrogation of the NICE SP&R Program leads to a clear conclusion. The issue is not the quality of NICE's analytical methods, the sophistication of its economic models, or its commitment

to evidence-informed decision making. The issue is more fundamental. The methodological framework responsible for improving NICE guidance demonstrates little recognition of the principles of representational measurement while continuing to endorse utilities, QALYs, and reference-case simulation models as though their quantitative legitimacy were already established.

This pattern is the defining characteristic of measurement inversion. Arithmetic is permitted to precede measurement, numerical constructions are treated as quantitative evidence, and methodological refinement is directed toward improving analytical techniques rather than establishing the scientific status of the quantities being analyzed. The interrogation demonstrates that this position is not confined to isolated methodological preferences but is embedded within the methodological research program that underpins NICE guidance.

Because NICE occupies one of the most influential positions in international health technology assessment, the implications extend well beyond the United Kingdom. NICE has contributed significantly to the global institutionalization of the reference-case paradigm. Consequently, if the methodological foundations of NICE fail to satisfy the standards of representational measurement, the issue is no longer one of improving NICE methods. It becomes a question of whether the contemporary HTA paradigm itself can continue to claim scientific legitimacy.

Given the established standards for representational measurement, if the objective is to estimate the impact of a therapy on a specified attribute there are only two scientifically defensible forms of quantitative measurement. Manifest attributes, such as survival time, hospital admissions, or treatment persistence, require linear ratio measures. Latent attributes, such as pain, fatigue, physical functioning, or need fulfilment, require Rasch logit ratio measures constructed from instruments satisfying the Rasch model. No third category of quantitative measurement exists. This proposition receives almost no endorsement in the NICE Science Policy and Research interrogation ($p = 0.05$; logit = –2.50), despite representing the logical consequence of representational measurement.

The implication is immediate. The EQ-5D-3L and EQ-5D-5L are not measures of therapy impact. They are descriptive classification systems that, through a valuation algorithm, generate composite utility scores whose measurement properties have never been demonstrated. Consequently, they cannot provide the quantitative foundation required for utilities, QALYs, cost-effectiveness ratios, or reference-case simulation models. Once this is recognized, the reference-case framework ceases to represent an alternative scientific methodology; it becomes a numerical construction built upon quantities that fail the established requirements for lawful measurement.

The conclusion reached in this assessment is therefore uncompromising. The challenge facing NICE is not methodological refinement but after over 20 years methodological reconstruction. A scientific HTA framework must begin with measurement rather than arithmetic, distinguish manifest from latent attributes, employ lawful ratio measurement appropriate to each, and restrict quantitative claims to those that are prospectively specified, evaluable, replicable, and falsifiable. Until those conditions are met, the reference-case paradigm remains an assumption-driven framework rather than a quantitative science.

HTA RECONSTRUCTION PROGRAM

The reference-case paradigm has reached the end of its scientific life. For more than four decades, health technology assessment (HTA) has relied upon utilities, quality-adjusted life years (QALYs), cost-effectiveness analysis, and reference-case simulation models to support claims regarding therapy impact. Recent assessments of HTA knowledge bases across government agencies, research centers, professional organizations, journals, and university programs demonstrate that this paradigm was established without satisfying the requirements of representational measurement. Arithmetic has consistently been placed before measurement. The result is measurement inversion, reinforced by curriculum inversion, in which generations of researchers and practitioners have been trained in analytical techniques without first acquiring the scientific principles required to determine whether those techniques can support lawful quantitative claims.

The HTA Reconstruction Program has been developed by Maimon Research LLC in response to this paradigm failure [9]. Its objective is not to reform the existing framework but to replace it with one grounded in representational measurement and the standards of normal science. Beginning with the principle that measurement must precede arithmetic, the program develops the scientific foundations required for credible therapy assessment: attributes, the principal scales of measurement, lawful ratio measurement, manifest and latent attributes, Rasch logit ratio measurement, and protocol-driven claims that are evaluable, replicable, and capable of falsification.

Designed for faculty, graduate students, researchers, manufacturers, formulary committees, and HTA agencies, the program provides a structured pathway from assumption-driven modelling to scientific HTA. It is intended to establish the competencies required for a new generation of health technology assessment in which evidence is grounded in lawful measurement rather than numerical construction, and where therapy impact claims meet the same scientific standards expected throughout the quantitative sciences

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 Kingdom: Measurement inversion and paradigm failure in health technology assessment. Logit Working Paper No 586. https://maimonresearch.com/logit-working-paper-no-586-june-2026/

[2] United Kingdom: Curriculum inversion and paradigm failure in health technology assessment. Logit Working Paper No 585. https://maimonresearch.com/logit-working-paper-no-585-june-2026/

[3] The end of the Reference Case: Reconstructing HTA. Logit Working Paper No 345. https://maimonresearch.com/logit-working-paper-no-345-june-2026-2/

[4] Stevens S. On the Theory of Scales of Measurement. *Science*. 1946;103(2684):677-80

[5] Krantz D, Luce R, Suppes P, Tversky A. Foundations of Measurement Vol 1: Additive and Polynomial Representations. New York: Academic Press, 1971

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

[7] Popper K. The Logic of Scientific Discovery. London: Hutchinson; 1959

[8] Popper K. Objective Knowledge: An Evolutionary Approach. Revised Edition. Oxford: Clarendon Press; 1979.

[9] Maimon Research Transition Program https://maimonresearch.com/hta-reconstruction-program-and-fees/