MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: FROM MEASUREMENT INVERSION TO A GLOBAL EPISTEMIC STRAITJACKET

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

LOGIT WORKING PAPER No 587 AUGUST 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

The United Kingdom has played a major role in the development and international diffusion of contemporary health technology assessment (HTA). For more than four decades, utilities, quality-adjusted life years (QALYs), cost-effectiveness ratios and reference-case simulation modelling have become embedded in reimbursement decision-making and subsequently transmitted internationally as accepted HTA practice. This paper revisits an earlier large language model interrogation of five influential UK HTA research-center knowledge bases—CHE, CRD, Oxford, ScHARR and Newcastle—to examine a more fundamental question: whether their quantitative frameworks demonstrate recognition that measurement must precede arithmetic.

Twenty-four canonical statements were assessed using categorical endorsement probabilities. Statements designated TRUE represent requirements of representational measurement, including unidimensionality, scale properties, ratio measurement, admissible arithmetic, dimensional homogeneity and Rasch measurement. FALSE statements represent assumptions incompatible with these requirements but embedded within legacy HTA. The results reveal a strikingly consistent pattern across all five centers. Measurement preceding arithmetic, the requirement that multiplication requires ratio measurement, and the requirement that representational-measurement axioms govern arithmetic each receive endorsement probabilities of only 0.10. Rasch-related measurement requirements fall as low as 0.05. Conversely, propositions supporting the established quantitative architecture of legacy HTA receive endorsement probabilities commonly between 0.85 and 0.95.

This near mirror-image endorsement structure is interpreted as measurement inversion: the arithmetic and analytical machinery are strongly possessed while the measurement principles required to justify that machinery are weakly possessed. The result is not simply an omission of measurement theory but a reversal of scientific sequence in which the operations required by the reference case effectively determine the properties presumed for the numbers.

The implications extend beyond the United Kingdom. UK institutions and methods have been major contributors to the internationalization of legacy HTA. Measurement inversion consequently travelled with a framework increasingly accepted as international best practice. The resulting institutional reinforcement across universities, HTA agencies, industry, consulting, journals and professional practice has created a global epistemic straitjacket in which ubiquity risks being mistaken for scientific validation. Escaping that constraint requires restoration of the proper scientific sequence: measurement first, arithmetic second.

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 [1]. 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 [2].

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.

The purpose of the present study is to examine whether this pattern is evident within the knowledge bases of an earlier study of measurement inversion in 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 Note that this paper is built on an earlier one [3]

Collectively, these institutions represent the intellectual core of UK HTA. Their research has influenced NICE methods, economic evaluation, evidence synthesis, technology appraisal, and health policy both within the United Kingdom and internationally.

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 five research groups, but to determine whether the intellectual foundations of UK 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 scales. 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 [4]. 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 [5] . 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.

LARGE LANGUAGE MODEL INTERROGATION OF KNOWLEDGE BASES

A large language model (LLM) provides a means of interrogating a defined knowledge base to estimate the extent to which particular concepts, assumptions and methodological propositions are represented and reinforced within it. The purpose is not to ask what the LLM *believes*. An LLM has no beliefs. Nor is the objective to infer the personal views or knowledge of individual researchers. The objective is to use the model as an interrogation instrument for examining the conceptual content of a defined knowledge space.

For the present UK assessment, the relevant knowledge spaces are those associated with five major university-affiliated HTA and health-economics research environments: the Centre for Health Economics (CHE), Centre for Reviews and Dissemination (CRD), the Oxford health-economics and HTA environment, the School of Health and Related Research (ScHARR), and the Newcastle health-services and health-economics environment. The interrogation draws upon the publicly represented methodological and research environments associated with these centers, including publications, research programs, methodological statements, teaching and training materials, reports and the wider HTA discourse with which they are associated.

The purpose of interrogation is narrowly defined. It asks to what extent would the knowledge base represented by each environment endorse a specified topic statement as an integral component of its knowledge and methodological framework?

This concept of endorsement is critical. The interrogation is not a simple keyword search. A knowledge base may contain the word *measurement* repeatedly without demonstrating possession of representational measurement. Similarly, references to *outcome measurement*, *preference measurement*, *psychometrics*, *utility measurement* or *quality-of-life measurement* do not establish recognition of scale properties, admissible transformations, dimensional homogeneity or the requirement that multiplication and division depend upon appropriate ratio properties.

The LLM instead evaluates the broader conceptual footprint of the topic. It asks whether the knowledge base explicitly or implicitly recognizes, reinforces and applies the proposition represented by the statement. The resulting categorical endorsement probability therefore represents the estimated strength with which the knowledge base supports the topic statement as an integral element of its conceptual structure.

This distinction makes LLM interrogation particularly useful for identifying measurement inversion. The question is not simply whether measurement terminology can be located. It is whether the knowledge base behaves as though measurement must precede arithmetic. Does it recognize that scale properties determine admissible mathematical operations? Does it require ratio properties before multiplication and division? Does it distinguish numerical assignment from measurement? Does it recognize unidimensionality and dimensional homogeneity? Does it distinguish the measurement requirements of manifest and latent attributes?

The interrogation therefore provides more than These propositions can then be contrasted with statements reflecting the established analytical structure of legacy HTA: treatment of utilities as quantitative measures, construction of QALYs, calculation of cost-per-QALY ICERs, use of preference scores, and reliance upon reference-case simulation models.

an inventory of topics. It estimates the direction and strength of conceptual endorsement across two competing scientific structures: one in which measurement determines permissible arithmetic and another in which the arithmetic required by the HTA framework proceeds without prior demonstration of the necessary measurement properties. Measurement inversion is revealed when these two patterns diverge systematically.

CATEGORICAL ENDORSEMENT PROBABILITIES

The diagnostic statements used in the interrogation are specified in advance. Each expresses a proposition relevant either to the requirements of representational measurement or to the assumptions and practices characteristic of legacy HTA. Under the representational-measurement framework, each statement can therefore be classified independently as TRUE or FALSE.

The task presented to the LLM is not to determine whether a statement *should* be accepted. Rather, it is to estimate the probability that the defined UK HTA knowledge base would endorse, imply or reinforce the statement as an integral component of its existing knowledge structur**e**, whether that statement is scientifically true or false under the measurement standard. This distinction is fundamental to interpretation.

A high endorsement probability for a TRUE statement indicates that the knowledge base strongly possesses or reinforces the relevant measurement principle. A low endorsement probability indicates that the principle has little conceptual presence within the knowledge base. Conversely, a high endorsement probability for a FALSE statement indicates that the knowledge base behaves as though the proposition were valid through its methods, assumptions, teaching or repeated application. It does **not** imply that individual researchers consciously believe something known to be false.

The probabilities are categorical rather than statistical estimates. They do not represent the percentage of faculty members who would agree with a statement, the proportion of publications containing particular terminology, or frequencies estimated from a sampled population. An LLM does not operate as a survey of individual respondents. It represents a domain through weighted linguistic and conceptual relationships. The endorsement probability therefore characterizes the strength of conceptual reinforcement within the knowledge base, not the prevalence of an opinion among individuals.

Categorical values are used deliberately to avoid false precision. Values such as 0.05, 0.10, 0.20, 0.50, 0.75, 0.80 and 0.85 represent progressively stronger degrees of conceptual possession or reinforcement. At the lower extreme, a probability such as 0.05 indicates near-absence of endorsement. At the upper end, values of 0.80 or 0.85 indicate that the proposition is routinely reinforced within the knowledge environment. Intermediate values represent weaker, inconsistent or ambiguous conceptual possession.

Three features of the knowledge base principally inform these estimates. First is the structural content of HTA discourse: the methods and assumptions repeatedly employed within the environment. Second is the conceptual visibility of measurement principles such as unidimensionality, scale integrity, dimensional homogeneity and admissible arithmetic. Third is the stability and consistency of the knowledge structure: whether the proposition is routinely reinforced, weakly represented or contradicted by established practice.

This allows measurement inversion to be represented directly. Consider two contrasting statements:

TRUE: *Measurement properties must be established before arithmetic operations are undertaken.*

FALSE: *A utility score bounded by 0 and 1 may be treated as possessing the quantitative properties required for multiplication by time.*

If the UK knowledge base assigns weak endorsement to the first while strongly endorsing, explicitly or through established practice, the second, the issue is not simply the absence of a topic called *representational measurement*. It demonstrates a reversal in the conceptual ordering of quantitative analysis. The requirements of measurement are weakly possessed while the arithmetic dependent upon the missing properties is strongly reinforced. That is the empirical signature of measurement inversion.

The same logic applies across the complete set of diagnostic statements. TRUE statements test possession of the scientific requirements governing measurement and admissible arithmetic. FALSE statements test the extent to which the knowledge base reinforces assumptions characteristic of a framework in which those requirements have been bypassed. The pattern across statements is consequently more informative than any individual probability.

The importance of the five-center UK interrogation therefore lies not primarily in the individual scores assigned to CHE, CRD, Oxford, ScHARR or Newcastle. It lies in the consistency of the endorsement structure across their knowledge bases. If centers with substantial expertise in health economics, HTA, evidence synthesis and decision modelling consistently show weak endorsement of TRUE statements concerning representational measurement while strongly endorsing FALSE statements implicit in legacy HTA practice, a common knowledge-base structure becomes visible.

The result can then be interpreted at the UK level: measurement inversion is not simply the omission of measurement theory from individual research programs. It is a systematic pattern in which the knowledge base strongly possesses the arithmetic and analytical machinery of HTA while weakly possessing the scientific measurement principles required to justify that machinery.

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

ENDORSEMENT RESULTS

Table 1 presents the results of the five UK research-center interrogations as categorical endorsement probabilities for each of the 24 canonical statements. Each statement was designated as either true or false according to the standards of representational measurement, scale theory, admissible arithmetic, and Rasch measurement.

TABLE 1: ITEM STATEMENT: RESPONSE AND ENDORSEMENT

STATEMENT	**RESPONSE 1=TRUE 0=FALSE**	**ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY**				
		CHE	**CRD**	**OXFORD**	**ScHARR**	**NEWCASTLE**
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	0.25	0.20	0.20	0.25
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	0.15	0.15	0.15	0.15
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	0.10	0.10	0.10	0.10
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.90	0.85	0.85	0.90	0.85
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.95	0.90	0.90	0.95	0.90

EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	0.90	0.90	0.90	0.90
THE QALY IS A RATIO MEASURE	0	0.85	0.90	0.90	0.95	0.90
TIME IS A RATIO MEASURE	1	0.95	0.95	0.95	0.95	0.95
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	0.10	0.10	0.10	0.10
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	0.85	0.85	0.90	0.85
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	0.10	0.10	0.10	0.10
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	0.05	0.05	0.05	0.05
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	0.05	0.05	0.05	0.05
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	0.85	0.85	0.90	0.85
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.95	0.90	0.90	0.95	0.90
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	0.15	0.15	0.15	0.15
QALYS CAN BE AGGREGATED	0	0.95	0.90	0.90	0.95	0.90

NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.65	0.80	0.80	0.70	0.70
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.95	0.90	0.90	0.95	0.90
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	0.60	0.65	0.65	0.60
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	0.05	0.05	0.05	0.05
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.35	0.35	0.35	0.35	0.35
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	0.20	0.25	0.20	0.20
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	0.05	0.05	0.05	0.05

Note: CHE Centre for Health Economics, York; CRD Centre for Reviews and Dissemination; Health Economics Research Centre/Oxford Population Health HTA Group (HERC); School of Health and Related Research (ScHARR), University of Sheffield; Institute of Health and Society. Newcastle HTA Group

The five-center UK interrogation presents an unusually coherent picture of. The principal finding is not that individual topics in representational measurement receive weak endorsement. It is that two opposed bodies of knowledge are endorsed in almost mirror-image fashion. Statements expressing the requirements that should govern lawful quantitative analysis receive consistently low endorsement probabilities, while statements representing assumptions embedded in legacy HTA receive consistently high endorsement probabilities. Across CHE, CRD, Oxford, ScHARR and Newcastle, the pattern is remarkably stable.

This matters because the five environments are not peripheral to UK HTA. They represent major concentrations of expertise in health economics, economic evaluation, evidence synthesis, modelling and technology assessment. The result therefore cannot plausibly be dismissed as the product of quantitatively unsophisticated knowledge bases. The interrogation instead identifies a much more serious problem: sophisticated analytical capability coexists with weak possession of the measurement principles required to determine whether that analysis is quantitatively admissible.

The clearest evidence lies in the propositions that should logically precede all quantitative HTA. **“**Measurement precedes arithmetic” receives an endorsement probability of only 0.10 at every center. “Meeting the axioms of representational measurement is required for arithmetic” likewise receives 0.10 throughout. “Multiplication requires a ratio measure” also receives 0.10 across all five centers. These are not specialist propositions at the margins of economic evaluation. They concern the conditions under which arithmetic itself has empirical meaning.

The consistency is striking. There is no center in which these principles emerge strongly enough to suggest a competing measurement tradition within UK HTA. CHE does not differ materially from ScHARR; Oxford does not provide a counterexample to York; Newcastle does not display a substantially different measurement foundation. The pattern is collective. On the LLM interrogation, the knowledge bases behave as though the scientific sequence measurement → scale → admissible arithmetic has little presence as an organizing constraint.

The result for unidimensionality reinforces the same conclusion. “Measures must be unidimensional” receives only 0.15 across all five centers. Yet without a defined attribute there is no coherent quantity to measure. This is particularly consequential for HTA because preference-based health-state systems routinely combine multiple dimensions of health into a single numerical index. Weak endorsement of unidimensionality means that one of the basic questions of measurement—*what single attribute does this number represent?*—has little conceptual force within the knowledge base.

The Rasch and latent-measurement results are still more extreme. The propositions concerning Rasch measurement as the pathway for latent attributes, the transformation of subjective observations under Rasch requirements, and the relationship of Rasch measurement to representational measurement receive probabilities as low as 0.05 at every center. The proposition that the outcome of interest for a latent attribute is possession of that attribute reaches only 0.20–0.25. On the stated interrogation framework, this is close to conceptual absence.

The statement “The QALY is a ratio measure”, designated false under the measurement framework used in the interrogation, receives endorsement probabilities of 0.85 to 0.95. The proposition that the QALY is dimensionally homogeneous receives 0.90 to 0.95. The proposition that QALYs can be aggregated receives the same range. EQ-5D preference algorithms producing interval measures receives 0.90 at every center. Summated subjective responses and Likert scores being treated as though they generate ratio measures receive 0.85 to 0.90.

This is the signature of measurement inversion in almost ideal form. The knowledge bases strongly reinforce the downstream numerical constructs while weakly reinforcing the upstream scientific conditions required to justify them. Put differently, the arithmetic is possessed; the measurement is not.

The QALY provides the most revealing example. The knowledge bases strongly recognize that time is a ratio measure: endorsement is 0.95 across all five centers. Yet endorsement of the equally necessary proposition that multiplication requires ratio measurement is only 0.10. At the same time, the QALY itself is strongly treated as though it possesses ratio and dimensional properties.

That combination is extraordinarily revealing. It suggests that the knowledge base recognizes the quantitative status of one component time without applying the same measurement test to the health-state value multiplied by it. The operation required by the QALY appears to have been accepted first, with the necessary properties of the other operand effectively presumed. That is precisely the reversal the concept of measurement inversion is intended to capture: the required arithmetic determines the presumed properties of the number rather than demonstrated measurement properties determining the admissible arithmetic.

The same failure then propagates downstream. Once the QALY is treated as quantitatively legitimate, aggregation becomes accepted; differences in QALYs become accepted; those differences enter ICER denominators; and increasingly elaborate simulation models manipulate the resulting constructs. Yet nothing later in that chain can repair a failure occurring at its beginning. If the QALY does not possess the properties required for its construction, aggregation cannot confer them. Subtraction cannot confer them. Division cannot confer them. Simulation cannot confer them. Arithmetic cannot rescue measurement failure by adding more arithmetic.

The false-statement endorsement pattern shows how completely this structure has become normalized. The proposition that ratio measures can have negative values receives 0.90–0.95 endorsement. Time trade-off preferences being treated as unidimensional receives 0.85–0.90. Reference-case simulations generating falsifiable claims receives 0.90–0.95. These are not isolated anomalies. They form a mutually reinforcing conceptual system in which the assumptions necessary for conventional HTA practice are strongly possessed even when the underlying measurement principles are weakly represented.

The falsification findings deserve particular attention. The true statement "Non-falsifiable claims should be rejected" actually performs considerably better than most representational-measurement items, with endorsement between 0.65 and 0.80. At first sight this might appear reassuring. But the false proposition "Reference-case simulations generate falsifiable claims" receives even stronger endorsement, between 0.90 and 0.95.

This exposes another form of inversion. The general principle of falsification is recognized, but the reference-case framework is nevertheless treated as satisfying it. The problem therefore appears not to be complete ignorance of the vocabulary of normal science. It is the failure to apply the principle critically to the dominant analytical structure. That distinction is important. A knowledge base can possess the language of science while failing to impose its constraints upon established practice.

The same is true of measurement. UK HTA can contain extensive references to *measurement*, *health outcomes*, *preferences*, *utilities* and *quality of life* without possessing the representational framework necessary to distinguish a numerical assignment from a measure. The interrogation is valuable precisely because it moves beyond terminology to the conceptual consequences of endorsement.

The near-identical probabilities across the five centers are also difficult to ignore. If the problem were confined to one research group, it could be attributed to local intellectual history or a particular methodological specialization. If one center strongly endorsed measurement-before-arithmetic while another did not, the UK knowledge base might reasonably be described as heterogeneous. That is not what the interrogation reports.

Instead, the endorsement profiles are almost interchangeable. The lowest probabilities repeatedly occur for the same measurement propositions; the highest probabilities repeatedly occur for the same legacy assumptions. This suggests not five independent methodological omissions but a shared epistemic architecture.

That architecture has profound historical significance because these centers have helped educate generations of health economists and HTA researchers. Knowledge bases reproduce themselves. Faculty teach what their discipline recognizes as important; students learn those priorities; graduates move into academia, agencies, consulting and industry; some become faculty; and the same analytical assumptions are transmitted again. Once a framework becomes institutionally mature, familiarity and ubiquity can easily substitute for foundational scrutiny. The interrogation therefore points beyond measurement inversion to epistemic reproduction.

There is a further point that makes the results particularly uncomfortable. The high endorsement of legacy HTA propositions cannot be defended on the grounds that the necessary measurement questions are simply irrelevant to economic evaluation. The reference case itself makes them unavoidable. QALYs require multiplication. ICERs require division. Preference measures are treated numerically. Model outputs are aggregated, averaged and compared. The framework is intrinsically dependent upon arithmetic. A discipline so dependent upon mathematical operations cannot coherently claim exemption from the scientific rules determining when those operations are admissible.

The issue is consequently not that UK HTA forgot to include an interesting chapter on measurement theory. It omitted the framework that should have governed the arithmetic upon which the discipline was built. This distinction separates a curricular gap from a foundational failure.

The results also undermine any defense based upon methodological sophistication. CHE, CRD, Oxford, ScHARR and Newcastle possess sophisticated expertise in modelling, evidence synthesis, economic evaluation and quantitative analysis. But greater sophistication downstream does not compensate for a missing measurement foundation upstream. Indeed, the more elaborate the subsequent analysis becomes, the more consequential the initial failure may become because inadmissible quantities are propagated through larger analytical structures. Simulation gives numerical assumptions computational reach. It does not give them measurement legitimacy.

The five-center evidence should therefore be interpreted as demonstrating three levels of measurement inversion. First is conceptual inversion: the principles of representational measurement receive weak endorsement while the numerical assumptions supporting legacy HTA receive strong endorsement. Second is analytical inversion: arithmetic operations are accepted before the properties required to justify those operations have been established. Third is institutional inversion: major academic environments reproduce and transmit the analytical framework without the measurement knowledge required to interrogate its foundations.

Taken together, these findings make it difficult to describe the problem as an isolated weakness of the QALY or a technical disagreement over health-state valuation. The interrogation suggests something much broader: the UK HTA knowledge base has developed around an analytical framework whose quantitative operations are strongly possessed while the scientific measurement principles required to justify them are weakly possessed or virtually absent.

That is why the endorsement probabilities matter. A simple absence-of-content analysis could show that representational measurement is rarely discussed. The categorical endorsement interrogation goes further. It demonstrates that the knowledge space does not merely leave a vacuum where lawful measurement should be. The vacuum is occupied by a strongly reinforced alternative set of assumptions that allows legacy HTA arithmetic to proceed as though the missing measurement properties had already been established**.** This is the defining epistemic feature of measurement inversion.

The implication for UK HTA is therefore severe. These five major academic environments did not simply develop an extensive analytical apparatus alongside an underdeveloped interest in measurement theory. On the interrogation presented here, they collectively display a knowledge structure in which the scientific order of quantitative inquiry has been reversed. Measurement should have determined the arithmetic. Instead, the arithmetic of the reference case appears to have determined what had to be presumed about measurement. For a putative quantitative discipline, it is difficult to imagine a more fundamental inversion.

CONCLUSION

The significance of the five-center UK interrogation extends well beyond the individual knowledge bases of CHE, CRD, Oxford, ScHARR and Newcastle. The United Kingdom has occupied a distinctive position in the development and international diffusion of contemporary health technology assessment. Through the development of health-economic methodology, the institutional influence of NICE, the prominence of UK academic research centers, and the international adoption of QALYs, cost-effectiveness thresholds and reference-case modelling, the UK has been one of the principal intellectual and methodological exporters of legacy HTA. What developed within the UK did not remain within the UK.

This gives the finding of measurement inversion particular importance. The interrogation suggests that some of the academic environments most closely associated with the development, application and transmission of modern HTA strongly possessed its analytical machinery while weakly possessing the representational-measurement principles required to establish whether that machinery was scientifically admissible. Measurement did not demonstrably provide the prior constraint upon arithmetic. Instead, utilities, QALYs, ICERs and simulation modelling became established components of the framework while their required measurement properties were assumed rather than independently demonstrated.

The consequence was not simply a British methodological error. The UK helped internationalize the framework in which the error was embedded.

As UK methods, academic expertise, guidelines and reference-case conventions acquired international authority, other jurisdictions had powerful incentives to adopt and adapt them. Universities trained students in the methods increasingly expected by HTA agencies. Manufacturers developed global economic-evaluation capabilities around them. Consultants reproduced them. Journals published them. Professional societies incorporated them into methodological discourse. National HTA agencies established their own reference cases. What began as a particular approach to informing reimbursement decisions progressively acquired the status of the accepted scientific language of HTA. Measurement inversion could therefore travel internationally without being recognized as measurement inversion. It travelled as best practice.

This history raises a fundamental question about the scientific status of legacy HTA. If the foundational quantities do not possess the measurement properties required for the arithmetic performed upon them, methodological maturity cannot repair the problem. Forty years of application cannot confer ratio properties upon a non-ratio quantity. Thousands of published cost-effectiveness analyses cannot establish the dimensional homogeneity of the QALY. Increasingly sophisticated probabilistic sensitivity analyses cannot make an inadmissible multiplication admissible. Microsimulation cannot transform a numerical construction into a measure. Institutional acceptance cannot substitute for scientific validation.

The challenge therefore reaches much further than criticism of a particular utility instrument, QALY algorithm or modelling technique. If measurement inversion is foundational to the reference case, then the scientific status of the reference case itself must be reconsidered.

The five-center findings also help explain why such reconsideration is difficult. Over several decades, legacy HTA has developed what can reasonably be described as an epistemic straitjacket. Its assumptions have become embedded simultaneously in academic research, graduate education, HTA agency requirements, reimbursement submissions, consulting practice, pharmaceutical-industry capabilities, journals and professional expectations. Each component reinforces the others. Agencies require the methods that universities teach; universities teach methods demanded by agencies and employers; manufacturers and consultants develop expertise to satisfy those requirements; journals publish analyses conducted within the accepted framework; and the resulting literature further legitimizes the framework.

The system thereby becomes self-validating. What is accepted is taught because it is accepted, applied because it is taught, required because it is routinely applied, and regarded as validated because it is universally required.

That is the epistemic straitjacket. Its strength lies not in the demonstration that its measurement foundations are sound, but in the extraordinary institutional density with which its assumptions have become embedded. Once an entire professional environment shares the same analytical vocabulary, fundamental questions become increasingly difficult to ask. Utilities are *measured*. QALYs are *calculated*. ICERs are *estimated*. Models are *validated*. Cost-effectiveness is *demonstrated*. The terminology itself can conceal the logically prior question: what has actually been measured?

The global reach of this straitjacket magnifies the problem. The measurement inversion identified in the early UK interrogations subsequently appears in assessments across different countries, agencies, research centers and educational environments. Different healthcare systems may employ different thresholds, decision rules, modelling conventions and reimbursement procedures, but the underlying quantitative architecture remains strikingly familiar. The same utilities, QALYs, ICERs and simulation structures recur, while the same prior requirements of representational measurement remain weakly represented.

This consistency should no longer be interpreted as independent confirmation of the scientific validity of legacy HTA. Global agreement can also represent global inheritance of the same unexamined assumptions**.** Widespread adoption is evidence of diffusion, not validation.

The historical role of the UK therefore carries an important contemporary implication. The intellectual influence that helped establish legacy HTA internationally can also contribute to its critical reassessment. UK universities and research centers are not permanently bound to the framework they helped develop. Nor should recognition of measurement inversion be interpreted principally as an accusation directed at the generations of researchers who worked within the accepted paradigm. The more consequential question is prospective: what happens once the inversion has been identified?

The scientific response should be straightforward. Return to the beginning. Define the attribute. Establish the required measurement structure. Determine the scale achieved. Identify the arithmetic permitted by that scale. Require dimensional homogeneity. Distinguish manifest from latent attributes. Only then construct quantitative claims and expose them to empirical evaluation, replication and potential falsification. The five UK knowledge bases examined here were among the early evidence that this sequence had been reversed. Their importance now lies less in the individual center results than in what they reveal about the intellectual architecture of a framework subsequently reproduced internationally. The UK helped give legacy HTA its global reach. The interrogation suggests that measurement inversion travelled with it.

The resulting epistemic straitjacket has constrained HTA for decades by defining which questions are asked, which methods are taught, which analyses are required and which numerical outputs are accepted as evidence. Escaping that constraint requires more than modifying the reference case. It requires restoring the scientific order that should have governed quantitative therapy evaluation from the outset: measurement first, arithmetic second**.**

Until that reversal is corrected, the central question facing legacy HTA remains unresolved: if its arithmetic was constructed without first establishing the measurement properties required to make that arithmetic admissible, on what scientific basis can its quantitative claims continue to be defended?

ACKNOWLEDGEMENT

I acknowledge that I have used OpenAI technologies, including the large language model, to assist in the development of this work. All final decisions, interpretations, and responsibilities for the content rest solely with me.

REFERENCES

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

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

[3] Langley P. 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/

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