MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: MEASUREMENT INVERSION AND PARADIGM FAILURE IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 586 JUNE 2026

www.maimonresearch.com

Tucson AZ

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 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 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. 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 [i]. 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 [ii] . 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 [iii]. 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 HTA knowledge space is defined pragmatically and operationally. For each jurisdiction, organization, or journal, the corpus consists of:

- published HTA guidelines
- agency decision frameworks
- cost-effectiveness reference cases
- academic journals and textbooks associated with HTA
- modelling templates, technical reports, and task-force recommendations
- teaching materials, methodological articles, and institutional white papers

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The "knowledge base" is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

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

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.

The interpretation of the results is straightforward. TRUE statements represent foundational principles of measurement science and the conditions required for valid quantitative claims. These include the propositions that measurement must precede arithmetic, that measures must be unidimensional, that different scales of measurement possess different analytical properties, that multiplication requires ratio measurement, that dimensional homogeneity is necessary for valid arithmetic operations, and that latent attributes require Rasch measurement if they are to support quantitative claims.

Endorsement probabilities for these TRUE statements are consistently low across all five research centers. This finding is important because low endorsement does not simply indicate disagreement. Rather, it indicates that these principles occupy little visible place within the HTA knowledge base.

The concepts are rarely articulated, rarely discussed, and rarely applied as governing constraints on quantitative analysis. In practical terms, the literature provides little evidence that the foundational requirements of measurement science are systematically recognized in the development, evaluation, or interpretation of HTA claims.

FALSE statements serve a different purpose. These statements represent assumptions that are incompatible with the standards of representational measurement but which have become embedded within the contemporary reference-case paradigm. Examples include the treatment of

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	1	0.10	0.10	0.10	0.10	0.10

MEASUREMENT IS REQUIRED FOR ARITHMETIC						
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	0	0.35	0.35	0.35	0.35	0.35

CAN ALWAYS BE COMBINED WITH A LOGIT SCALE						
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 (IHS)

---

utility scores as ratio measures, the assumption that QALYs possess ratio properties, the belief that summated ordinal responses can generate ratio scales, the acceptance of dimensional incoherence, the aggregation of QALYs, and the assumption that reference-case simulation models generate empirically testable evidence.

For these FALSE statements, endorsement probabilities are consistently high. This does not imply that the literature explicitly states the false proposition. Rather, it indicates that the knowledge base behaves as though the proposition were true. The proposition is embedded in analytical practice, methodological guidance, model construction, evidence synthesis, and economic evaluation. High endorsement therefore reflects institutional acceptance of assumptions that conflict with the requirements of measurement science.

The combined pattern is the defining characteristic of measurement inversion. Statements that should be strongly endorsed receive consistently low probabilities, while statements that should be rejected receive consistently high probabilities. The issue is therefore not merely the absence of measurement knowledge. More importantly, the knowledge base systematically reinforces assumptions that are incompatible with the principles of representational measurement. Arithmetic is accepted before measurement is established, and quantitative claims are advanced without first demonstrating that the quantities involved satisfy the conditions necessary to support those claims.

The significance of this reversal cannot be overstated. It provides direct evidence that the contemporary HTA framework is not simply detached from measurement science but operates according to assumptions fundamentally opposed to it. The endorsement pattern therefore points beyond isolated methodological weaknesses to a broader phenomenon of measurement inversion embedded within the intellectual foundations of the reference-case paradigm.

ENDORSEMENT RESULTS: CENTRE FOR HEALTH ECONOMICS

The interrogation of the Centre for Health Economics knowledge base occupies a unique position within the present assessment of UK health technology assessment. Unlike the other institutions examined, CHE is not simply a participant in the HTA environment. It is one of the principal intellectual architects of the contemporary reference-case paradigm. Through its research programs, methodological contributions, training activities, and longstanding influence on NICE methods, CHE has played a major role in shaping the framework that now dominates health technology assessment not only in the United Kingdom but internationally. Consequently, the interrogation of CHE is more than an assessment of a university research center. It is an assessment of the intellectual foundations of the reference-case paradigm itself.

The results reveal a striking and internally consistent pattern. Propositions grounded in representational measurement, ratio measurement, unidimensionality, latent attribute measurement, and Rasch theory receive very low endorsement probabilities. At the same time, propositions embedded in the utility-QALY framework receive very high endorsement probabilities. The result is a classic example of measurement inversion: a knowledge base that strongly supports the outputs of quantitative analysis while giving little recognition to the conditions required to justify those outputs.

The most revealing findings concern the relationship between measurement and arithmetic. The proposition that multiplication requires a ratio measure receives a probability of only 0.10. The proposition that measurement precedes arithmetic also receives only 0.10. Similarly, the proposition that the axioms of representational measurement are required before arithmetic operations can be applied receives a probability of 0.10. These are among the lowest endorsement probabilities observed in the interrogation.

Their importance cannot be overstated. The contemporary reference-case framework is built almost entirely upon arithmetic operations. Utility scores are multiplied by time to generate QALYs. QALYs are aggregated across individuals and populations. Costs are divided by QALYs to generate cost-effectiveness ratios. Simulation models extend these calculations over decades and hypothetical populations. Every stage of the analytical framework depends upon arithmetic. Yet the interrogation indicates little recognition that arithmetic itself requires prior demonstration of measurement properties.

This finding strikes directly at the intellectual foundations of the reference case. The framework assumes that utility values possess the properties necessary to support multiplication. It assumes that QALYs possess the properties necessary to support aggregation. It assumes that cost-per-QALY ratios possess the properties necessary to support comparison and decision making. However, the propositions required to establish these assumptions receive consistently low endorsement. The implication is unavoidable: the arithmetic is accepted while the measurement foundations are assumed.

The treatment of utility measurement illustrates this problem particularly clearly. The interrogation reveals strong endorsement of propositions embedded within the utility-QALY framework. Time trade-off preferences are effectively treated as suitable inputs to utility construction. EQ-5D

preference algorithms are accepted as generating quantities suitable for economic evaluation. QALYs are treated as legitimate quantitative outcomes. Yet the propositions necessary to justify these interpretations receive little support. Multiplication requires ratio measurement receives only 0.10. Representational measurement receives only 0.10. Unidimensionality receives only 0.15.

The contradiction is fundamental. Utility instruments are treated as though they generate quantities suitable for arithmetic manipulation, while the conditions necessary to establish those properties are largely absent from the knowledge base. The result is a framework in which utility values acquire legitimacy through use rather than through demonstration of measurement status.

The same pattern is evident in the treatment of QALYs. The proposition that the QALY is a ratio measure is strongly endorsed within the logic of the reference-case framework. Similarly, the proposition that QALYs can be aggregated receives a very high endorsement probability. Yet both claims depend upon measurement properties that are not independently established. If the utility component lacks ratio properties, multiplication by time is inadmissible. If the resulting QALY lacks dimensional homogeneity, aggregation becomes equally problematic. The interrogation suggests little recognition of these issues.

This is the essence of measurement inversion. Rather than asking whether a quantity possesses the properties required for a particular operation, the operation is performed and the resulting quantity is assumed to be meaningful. Arithmetic becomes the source of legitimacy. Measurement is treated as an afterthought.

The low endorsement of unidimensionality further reinforces this conclusion. Measurement requires that an attribute represent a single dimension. Without unidimensionality there may be numerical assignment, ranking, or aggregation, but there is no measurement. Yet the proposition that measures must be unidimensional receives only 0.15. This is especially significant because utility systems combine multiple domains of health into a single numerical output. Mobility, pain, anxiety, self-care, and other dimensions are collapsed into a single value. Without demonstrating unidimensionality, there is no basis for claiming that the resulting quantity represents a measure of a single attribute.

The latent attribute findings are equally important. The propositions concerning Rasch measurement receive the lowest endorsement probabilities in the interrogation. The proposition that there are only two classes of measurement—linear ratio measurement for manifest attributes and Rasch logit ratio measurement for latent attributes—receives only 0.05. The proposition that transformation of subjective observations into measurement is only possible through Rasch rules receives 0.05. The proposition that the Rasch logit ratio scale provides the basis for assessing therapy impact in latent traits also receives 0.05.

These findings indicate that the conceptual framework required for latent attribute measurement occupies almost no visible place within the CHE knowledge base. This omission is particularly striking because many of the outcomes central to contemporary HTA are latent attributes. Quality of life, fatigue, pain, treatment satisfaction, functioning, and psychological wellbeing are not directly observable. They require a measurement model capable of estimating possession of the

attribute. Yet the interrogation suggests that this measurement problem is largely invisible within the intellectual environment responsible for constructing the utility-QALY framework.

The absence of Rasch measurement has far-reaching implications. Once Rasch disappears, latent outcomes are left without a scientifically defensible measurement framework. Ordinal responses can be summed, weighted, transformed, and incorporated into utility algorithms without first establishing whether they support measurement. The resulting quantities may be useful as descriptive constructs, but they cannot be assumed to support arithmetic operations simply because they are numerical.

The proposition that the outcome of interest for latent traits is possession of the trait receives only 0.25. This finding is particularly revealing. Measurement of latent attributes is fundamentally concerned with estimating possession of the attribute. Therapy impact is assessed by changes in possession. Without recognition of this principle, latent outcomes become detached from measurement theory and are instead treated as numerical inputs to broader evaluative frameworks.

The implications extend directly to simulation modelling. The interrogation reveals strong endorsement of the proposition that reference-case simulations generate meaningful outputs. Yet simulation models do not create measurement. They manipulate quantities already assumed to be valid. If the utility scores entering a model lack ratio properties, every QALY produced by that model inherits the same defect. Increasing sophistication cannot compensate for defective inputs. A model cannot create measurement where measurement does not exist.

This observation is particularly important for CHE because the center\ has been one of the principal contributors to the development of decision modelling within HTA. The interrogation suggests that modelling occupies a far more prominent position within the knowledge base than measurement itself. The result is that the analytical machinery of the reference case receives extensive development while the measurement status of the quantities entering that machinery receives comparatively little attention.

The broader significance of the findings lies in CHE's influence on the international HTA community. Through NICE and through the dissemination of reference-case methods, concepts developed and promoted within CHE have shaped HTA practice across multiple jurisdictions. The interrogation therefore provides insight not only into a single institution but into the intellectual foundations of the wider HTA paradigm.

The consistency of the profile is particularly revealing. The same propositions receive low endorsement. The same assumptions receive high endorsement. The same absence of measurement science recurs throughout the knowledge base. This is not evidence of isolated omissions. It is evidence of a coherent intellectual framework in which quantitative outputs are accepted while the conditions necessary to justify those outputs remain largely absent.

The consequence is that the reference case emerges not as a framework grounded in measurement but as a framework grounded in assumptions regarding measurement. Utility values are assumed to possess ratio properties. QALYs are assumed to be dimensionally homogeneous. Simulation

outputs are assumed to constitute evidence. These assumptions become embedded within teaching, research, and policy, eventually acquiring the appearance of established scientific fact.

This explains the persistence of the reference-case paradigm despite its measurement deficiencies. The concepts required to identify the problem are themselves largely absent from the knowledge base. Researchers are trained to construct models, estimate QALYs, and perform economic evaluations without first being trained to determine whether the quantities involved satisfy the requirements of measurement science. The analytical impossibility remains invisible because the intellectual tools required to recognize it are missing.

The CHE interrogation therefore leads to a stark conclusion. The knowledge base demonstrates strong endorsement of the reference-case paradigm while giving little recognition to the principles of representational measurement upon which that paradigm ultimately depends. The result is measurement inversion at the intellectual center of UK HTA. The significance extends beyond a single institution because CHE has played a major role in shaping the reference-case framework itself. If measurement inversion is present here, it reaches directly into the foundations of the paradigm.

For claims regarding therapeutic value and therapy impact, the implications are profound. A framework that depends upon arithmetic operations cannot exempt itself from the requirements governing arithmetic. If ratio measurement has not been demonstrated, multiplication is inadmissible. If dimensional homogeneity has not been established, aggregation is inadmissible. If latent attributes have not been measured, quantitative claims regarding their possession remain unsupported. These are not methodological preferences. They are requirements of quantitative science.

The interrogation therefore reinforces the broader conclusion emerging from the UK assessments. The challenge confronting HTA is not one of methodological refinement. It is not a matter of better utility instruments, improved modelling techniques, or more sophisticated simulations. The problem is foundational. The concepts required to establish measurement have been displaced by the outputs of arithmetic. CHE's knowledge base provides perhaps the clearest example of this inversion because it lies so close to the intellectual origins of the reference-case paradigm itself.

The conclusion is unavoidable. The CHE knowledge base demonstrates strong commitment to economic evaluation, utility assessment, QALY construction, and decision modelling while giving little recognition to the measurement science required to support those activities. The result is measurement inversion embedded within the intellectual foundations of contemporary HTA. For a paradigm that claims to generate robust estimates of therapy impact, this finding is devastating. The issue is no longer one of reform. It is one of paradigm failure.

ENDORSEMENT RESULTS: CENTRE FOR REVIEWS AND DISSEMINATION

The interrogation of the Centre for Reviews and Dissemination knowledge base presents a distinctive form of measurement inversion. CRD is not primarily identified with the theoretical construction of the reference case in the same way as the Centre for Health Economics, nor is it

principally an operational appraisal center in the same sense as ScHARR. Its authority rests on evidence synthesis: systematic review, technology assessment review, critical appraisal, dissemination, and the organization of research findings for policy and practice. For that reason, the CRD interrogation raises a particularly sharp question: what happens when evidence is synthesized before measurement has been established?

The profile shows strong recognition of evidence-facing principles. Time is endorsed as a ratio measure with a probability of 0.95. Non-falsifiable claims should be rejected receives a probability of 0.80. The logit as the natural logarithm of the odds ratio receives moderate endorsement at 0.60. These results suggest a knowledge base oriented toward empirical inquiry, assessment, and claims evaluation. CRD is clearly not indifferent to evidence. The difficulty is that evidence synthesis and measurement are not the same activity. A systematic review can assemble studies, evaluate their quality, assess bias, compare findings, and summarize conclusions. It cannot transform quantities that are not measures into measures.

This is the central significance of the CRD profile. The propositions that define the relationship between measurement and arithmetic receive very low endorsement. Multiplication requires a ratio measure receives only 0.10. Measurement precedes arithmetic receives only 0.10. Meeting the axioms of representational measurement is required for arithmetic also receives only 0.10. These are foundational propositions. Without them, there is no principled basis for deciding whether the numerical objects entering a review are capable of supporting the claims being synthesized.

The low endorsement of representational measurement is especially damaging. Representational measurement provides the framework that distinguishes measurement from numerical assignment. It asks whether numbers correspond to empirical attributes in ways that justify particular arithmetic operations. Without this framework, evidence synthesis risks treating all numerical outputs as if they were commensurable evidence. Utility values, preference scores, patient-reported outcome totals, QALYs, and cost-effectiveness ratios may be reviewed and summarized, but their measurement status remains unexamined.

This creates the evidence synthesis problem. A systematic review may be methodologically impeccable and still synthesize measurement garbage. Search strategies may be comprehensive. Inclusion criteria may be explicit. Risk of bias may be carefully assessed. Study selection may be reproducible. Data extraction may be transparent. None of this can answer the prior question: are the quantities being synthesized measures? If the answer is no, the review process does not cure the defect. It merely organizes it.

The distinction is decisive. Evidence synthesis can strengthen evidence only where the underlying evidence is grounded in valid measures. If the quantities entering the synthesis are ordinal scores treated as interval or ratio measures, or utility values treated as proportions, or QALYs treated as dimensionally homogeneous outcomes, then the synthesis inherits the same measurement failure. A meta-analysis of non-measures does not create measurement. A systematic review of QALY estimates does not establish that QALYs are valid measures. A synthesis of utility scores does not demonstrate that utility scores possess ratio properties.

The CRD profile indicates little recognition of this problem. Interval measures lack a true zero receives only 0.25. Measures must be unidimensional receives only 0.15. Multiplication requires ratio measurement receives only 0.10. These low probabilities suggest that the measurement properties of reported quantities are not treated as governing constraints. This is not a minor omission. It affects the entire logic of evidence synthesis in HTA because technology assessments often depend on quantities whose measurement status is questionable or unproven.

The treatment of QALYs and utilities illustrates the point. The propositions associated with the reference-case paradigm receive strong endorsement. The QALY is treated as a legitimate quantitative outcome. QALYs are treated as aggregable. Reference-case simulations are treated as capable of generating claims. Utility algorithms are treated as producing usable numerical inputs. Yet the concepts required to justify these assumptions—ratio measurement, dimensional homogeneity, representational measurement, unidimensionality—receive weak endorsement. The result is a knowledge base in which the products of the reference case are available for synthesis, but the measurement foundations of those products are largely invisible.

This is measurement inversion expressed through evidence synthesis. In economic evaluation, inversion appears when arithmetic is performed before measurement is established. In CRD's case, inversion appears when numerical claims are reviewed, summarized, and disseminated before the measurement status of those claims is established. The sequence is still reversed. Synthesis precedes measurement. The review process assumes the legitimacy of the quantities under review rather than requiring their measurement properties to be demonstrated.

The implications for latent attributes are particularly serious. HTA frequently relies on outcomes such as pain, fatigue, functioning, quality of life, treatment satisfaction, mental health, and patient experience. These are latent attributes. They cannot be directly observed. They require a measurement model capable of estimating possession of the attribute. Yet the CRD profile gives little recognition to latent attribute measurement. The proposition that the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits receives 0.05. The proposition that transforming subjective responses requires Rasch rules receives 0.05. The proposition that the outcome of interest for latent traits is possession of the trait receives only 0.20.

This omission matters because systematic reviews frequently encounter patient-reported outcome measures, quality-of-life instruments, preference-based scores, and utility values. Without Rasch measurement, ordinal responses remain ordinal responses. Summing them, weighting them, aggregating them, or converting them into utilities does not establish measurement. The CRD profile suggests that the knowledge base does not give this issue central importance. As a result, reviews may synthesize latent-outcome studies without first determining whether the latent attribute has been measured.

The absence of the manifest-latent distinction is equally revealing. Manifest attributes such as mortality, survival time, hospital admissions, adverse events, resource utilization, and treatment persistence may support direct observation and linear ratio measurement. Latent attributes require Rasch logit ratio measurement. These are different measurement problems. Yet the proposition that there are only two classes of measurement—linear ratio and Rasch logit ratio—receives 0.05. The proposition that manifest and latent measurement cannot simply be collapsed into a common

scale is weakly recognized. This means that fundamentally different types of quantities can enter evidence synthesis as if they occupied the same measurement space.

The result is not merely technical confusion. It has direct implications for the scientific status of review conclusions. A review that synthesizes survival time, hospital admissions, EQ-5D utilities, QALYs, quality-of-life scores, and modelled cost-effectiveness estimates must first ask whether these quantities are measures of the same kind, or measures at all. If this question is not asked, the review may appear rigorous while leaving the core measurement problem untouched.

This is why the CRD interrogation is so important. CRD represents the evidence-synthesis arm of the UK HTA system. If evidence synthesis proceeds without representational measurement, then it can reinforce the authority of the reference case by treating its numerical products as evidence. The review process may then become a mechanism through which measurement inversion is institutionalized. The stronger and more systematic the review process, the more authoritative the synthesis appears, even if the quantities being synthesized lack measurement status.

This is not a criticism of review discipline as such. Systematic review is a powerful scientific method when applied to valid evidence. The problem arises when the review process is applied to numerical outputs whose measurement properties have not been established. Under those conditions, rigor in synthesis can conceal weakness in measurement. The appearance of evidence strengthens while the foundations of that evidence remain defective.

The CRD profile also helps explain the persistence of the reference-case paradigm. If the institutions responsible for evidence synthesis do not place measurement science at the center of evaluation, then the reference case is protected from one of the most fundamental forms of scrutiny. Utilities are reviewed as inputs. QALYs are reviewed as outcomes. Simulation outputs are reviewed as evidence. But the prior question—whether these quantities satisfy the axioms of measurement—is not treated as decisive.

The broader UK implication is clear. CHE reveals economic evaluation before measurement. ScHARR reveals appraisal implementation before measurement. Oxford reveals empirical evaluation before measurement. Newcastle reveals applied HTA review before measurement. CRD reveals evidence synthesis before measurement. These are different institutional routes to the same paradigm failure. Across the system, the reference case is accepted while the measurement conditions required to justify it remain largely absent.

For claims of robust therapy impact, the CRD findings are devastating. Robustness cannot be achieved through synthesis alone. It cannot be achieved by combining more studies, improving review protocols, or increasing transparency if the underlying quantities are not measures. A robust synthesis of invalid quantities remains invalid. The problem is not the method of synthesis; it is the status of the material being synthesized.

The conclusion is therefore unavoidable. The CRD knowledge base demonstrates strong commitment to evidence synthesis and empirical review, but weak recognition of the measurement science necessary to determine whether the evidence being synthesized is valid quantitative evidence. This is measurement inversion in review form. It is not sufficient to ask what the

literature reports. It is necessary first to ask whether the quantities reported in the literature are measures. Until that question is placed at the center of HTA review, evidence synthesis risks becoming the systematic organization of numerical constructions rather than the synthesis of scientific evidence.

For CRD and for UK HTA more broadly, the challenge is not to make evidence synthesis more elaborate. The challenge is to place measurement scrutiny before synthesis. Every quantitative claim entering a review should be required to identify the target attribute, establish whether it is manifest or latent, demonstrate the relevant ratio measurement properties, and show that the arithmetic operations applied are admissible. Without that sequence, synthesis cannot support robust claims regarding therapy impact. It can only reproduce the measurement failures of the studies and models it reviews.

ENDORSEMENT RESULTS: OXFORD HTA AND HEALTH ECONOMICS

The interrogation of the Oxford HTA and Health Economics knowledge base reveals a distinctive form of measurement inversion. Oxford is strongly associated with empirical evaluation, health economics, outcomes research, trial-based analysis, population health, and policy-relevant evidence. This gives the Oxford profile a slightly different emphasis from CHE, CRD, ScHARR, or Newcastle. The knowledge base is likely to show strong commitment to empirical assessment, intervention comparison, health outcomes, and claims capable of evidentiary evaluation. Yet the central weakness remains the same: the concepts required to determine whether the quantities entering those evaluations are lawful measures occupy only a marginal position.

The profile shows strong endorsement where Oxford's empirical orientation would be expected to appear. Time as a ratio measure receives 0.95. Non-falsifiable claims should be rejected receives 0.80, the highest among the UK centers considered so far. The logit as the natural logarithm of the odds ratio receives moderate recognition at 0.65. These findings suggest a knowledge base comfortable with empirical inquiry, statistical reasoning, observed outcomes, and evaluation. This is not a knowledge base indifferent to evidence. The problem is more subtle and more serious. Evidence is emphasized, but measurement is not established as the prerequisite for quantitative evidence.

The central measurement propositions receive very low endorsement. Multiplication requires a ratio measure receives only 0.10. Measurement precedes arithmetic also receives 0.10. The proposition that meeting the axioms of representational measurement is required for arithmetic receives 0.10. These findings are decisive. They indicate that the knowledge base does not treat measurement as the governing condition for quantitative claims. Arithmetic appears to be accepted as part of economic evaluation and outcomes analysis without first requiring demonstration that the quantities involved possess the scale properties necessary to support the operations performed.

This is the essential structure of measurement inversion. Oxford may be strong on empirical evaluation, but empirical evaluation does not rescue a quantity whose measurement status has not been established. A study may be well designed, a trial may be large, an observational analysis may be carefully adjusted, and an economic evaluation may be technically sophisticated. None of

this establishes that utility scores possess ratio properties, that QALYs are dimensionally homogeneous, or that patient-reported outcomes have been transformed into valid measures of latent attribute possession. Measurement is prior to analysis. If this step is missing, the resulting quantitative claims rest on numerical construction rather than lawful measurement.

The low endorsement of representational measurement is particularly important. Representational measurement provides the formal framework for determining whether numerical assignments correspond to empirical attributes in ways that support admissible arithmetic. Without this framework, there is no principled distinction between a number and a measure. Utility scores, composite indices, quality-of-life values, and QALYs may appear quantitative because they are numerical, but numerical form is not measurement. The Oxford interrogation indicates that this distinction receives little recognition.

The same issue applies to scale theory. Interval measures lack a true zero receives only 0.20. Multiplication requires ratio measurement receives 0.10. Yet the QALY depends entirely on multiplication. Survival time may be a ratio measure, but the utility value used to weight that time must also possess ratio properties if multiplication is to be admissible. If it does not, the resulting QALY is not a measure of quality-adjusted survival. It is a numerical artefact created by applying arithmetic to quantities that do not support the operation. The interrogation suggests that this problem is not central to the Oxford knowledge base.

The treatment of QALYs and utilities confirms this conclusion. The propositions challenging the QALY and utility framework receive high endorsement in the direction of the reference-case paradigm. The QALY is a ratio measure receives 0.90 as an endorsed false proposition. The QALY is dimensionally homogeneous receives 0.90. QALYs can be aggregated receives 0.90. EQ-5D-3L preference algorithms create interval measures receives 0.90. These results indicate strong acceptance of the reference-case treatment of utility-based quantities, while the measurement standards required to justify that treatment remain weakly endorsed.

The contradiction is central. Oxford's knowledge base is oriented toward robust empirical analysis, but the reference-case quantities that enter such analysis remain unexamined in measurement terms. Robustness cannot be achieved merely through better data, larger samples, improved modelling, or more sophisticated econometrics. A robust estimate of therapy impact requires a valid measure of the attribute being assessed. If the quantity is not a measure, robustness attaches only to the manipulation of the numerical construction, not to the measurement of therapeutic impact.

The low endorsement of unidimensionality reinforces the problem. Measures must be unidimensional receives only 0.15. This is important because HTA routinely employs instruments that combine multiple domains into single utility values or composite scores. Mobility, pain, anxiety, usual activities, self-care, function, and wellbeing may be collapsed into a single index. Unless the attribute is demonstrated to be unidimensional, this is aggregation, not measurement. The numerical result may be useful for description or administrative comparison, but it cannot support quantitative claims unless the measurement model has demonstrated that a single attribute is being measured.

The latent attribute results are equally significant. The Rasch logit ratio scale as the basis for assessing therapy impact in latent traits receives only 0.05. Transformation of subjective responses through Rasch rules receives only 0.05. The outcome of interest for latent traits is possession of the trait receives only 0.25. These results indicate little recognition of the central measurement challenge posed by patient-centric outcomes. Pain, fatigue, anxiety, physical functioning, quality of life, treatment satisfaction, and need fulfilment are not directly observable. They are latent attributes. Their measurement requires a model capable of estimating possession of the attribute. Rasch measurement provides the framework for doing this. Without it, ordinal responses remain ordinal responses, whatever numerical operations are subsequently imposed upon them.

This omission is particularly damaging for a knowledge base concerned with outcomes research. Outcomes research is not measurement research unless it first establishes the measurement status of the outcome. A reported outcome may be clinically relevant, patient-centric, policy-relevant, or statistically significant, but these qualities do not make it a measure. If the outcome is latent, then the question becomes whether possession of the attribute has been measured through a valid measurement model. If this question is not asked, the analysis proceeds on assumption.

The Oxford profile therefore reveals a distinct version of measurement inversion: empirical evaluation before measurement. The knowledge base values evidence, claims assessment, comparative evaluation, and real-world relevance. Yet it gives little recognition to the measurement foundations required to make quantitative evidence meaningful. This differs from CHE, where the emphasis is economic modelling before measurement; from CRD, where the issue is synthesis before measurement; and from ScHARR, where the issue is appraisal implementation before measurement. In Oxford's case, the problem is empirical analysis before measurement.

This distinction matters because it shows that measurement inversion can survive even in knowledge environments that take evidence seriously. The problem is not anti-empiricism. It is the failure to distinguish evidence from measurement. Evidence concerning a non-measure does not become stronger simply because it is empirically collected. If the numerical representation of the attribute is invalid, then statistical analysis of that representation cannot produce valid quantitative claims. The problem lies not in the statistical machinery but in the measurement foundation.

This has direct implications for trial-based economic evaluation. A trial may observe survival, hospital admissions, adverse events, discontinuation, and resource use. These manifest attributes may support linear ratio measures if properly specified. But the same trial may also report utilities, quality-of-life scores, symptom scales, or patient-reported outcomes. These latent attributes require a different measurement framework. If both manifest and latent outcomes are treated as interchangeable numerical variables, the distinction between observation and measurement disappears. The proposition that there are only two classes of measurement—linear ratio and Rasch logit ratio—receives only 0.05. This indicates that the Oxford knowledge base gives little attention to the two-measure framework required for therapy impact claims.

The proposition that a linear ratio scale for manifest claims can always be combined with a logit scale receives a probability of 0.35. This indicates some recognition that scales cannot simply be combined at will, but the result remains weak. The distinction is crucial. Manifest and latent

attributes may both be relevant to therapy assessment, but they cannot be collapsed into a single artificial index without demonstrating the admissibility of the operation. A therapy may reduce hospital admissions and improve fatigue possession, but these are different attributes measured through different forms of ratio measurement. A lawful claims framework would present them separately, each with its own measurement model and falsification rule.

The reference-case paradigm avoids this discipline by converting diverse outcomes into utility weights and QALYs. This is precisely what measurement science challenges. Utility-based aggregation obscures the attribute structure of therapy impact. Instead of asking what attribute changed, how it was measured, and whether the claim is evaluable, the reference case generates a common numerical currency. The problem is that the common currency lacks demonstrated measurement status. The Oxford interrogation suggests that this problem is not given central recognition.

The broader significance is clear. Oxford is not a peripheral institution. It is one of the United Kingdom's most important environments for health economics, outcomes research, and population-health evaluation. If the Oxford knowledge base gives little recognition to representational measurement, ratio measurement, Rasch measurement, and the manifest-latent distinction, then the problem extends beyond specialist HTA modelling. It reaches the wider empirical research culture that supports HTA.

The conclusion is therefore not that Oxford lacks evidence orientation. The conclusion is that evidence orientation is insufficient. A scientific HTA framework requires evidence grounded in measurement. Without measurement, empirical analysis may generate precise estimates of quantities whose scientific status is unresolved. Such estimates may appear robust, but robustness is misplaced if the object being estimated is not a measure.

For claims regarding therapeutic impact, this is devastating. The reference case claims to support robust estimates of value, effectiveness, and cost-effectiveness. Yet its central constructs depend on utilities, QALYs, aggregation, and simulation outputs whose measurement properties remain unproven. The Oxford interrogation reinforces the broader UK finding: the concepts required to expose this weakness are largely absent from the knowledge base. The result is not merely a methodological gap. It is a structural failure in the intellectual foundations of HTA.

The Oxford HTA and Health Economics knowledge base therefore demonstrates measurement inversion in an empirical evaluation setting. Outcomes are emphasized, evidence is valued, and falsifiability receives relatively strong recognition. Yet the measurement standards required to make quantitative evidence scientifically meaningful receive low endorsement. The framework remains trapped in arithmetic before measurement. The consequence is the continued normalization of reference-case outputs despite their failure to satisfy ratio measurement standards. The challenge is not to refine the reference case, but to replace it with a framework in which attributes are specified, manifest and latent outcomes are distinguished, measurement precedes arithmetic, and claims regarding therapy impact are evaluable, replicable, and falsifiable.

ENDORSEMENT RESULTS: ScHARR

The ScHARR profile displays the signature pattern of measurement inversion. Scientifically correct propositions grounded in measurement theory receive low endorsement probabilities, while propositions embedded within the reference-case paradigm receive strong endorsement. This is not a random pattern of omission. It is a structured reversal of scientific priority. The knowledge base strongly supports the operational practices of contemporary HTA while giving little recognition to the measurement conditions required to justify those practices.

The most important findings concern the foundations of arithmetic. The proposition that multiplication requires a ratio measure receives a probability of only 0.10. The proposition that measurement precedes arithmetic also receives 0.10. The proposition that meeting the axioms of representational measurement is required for arithmetic receives 0.10. These are not specialist or marginal claims. They express the basic requirements for lawful quantitative analysis. If multiplication requires ratio measurement, and if the reference case depends upon multiplication of utility values by time, then the measurement status of the utility value becomes decisive. Yet the interrogation suggests that this requirement occupies almost no meaningful position within the ScHARR knowledge base.

This omission is particularly consequential because the NICE reference case is built around cost-utility analysis and the use of QALYs. The NICE methods guide states that economic evaluations should include analyses using the reference-case methods, and the reference case is intended to support consistency across health technology evaluations. ScHARR operates within this environment. Its contribution to evidence review, appraisal support, and economic modelling is therefore linked to the same framework whose central quantities require scrutiny under measurement theory. If the reference case requires utilities and QALYs, then the question is unavoidable: do these quantities possess the measurement properties necessary to support the arithmetic imposed upon them?

The interrogation indicates that this question is not given central attention. The QALY is treated as if it were a legitimate quantitative construct, while the conditions necessary to establish that legitimacy are weakly endorsed or effectively absent. The proposition that the QALY is a ratio measure is marked as false, yet the endorsement probability of the response indicates strong support for the reference-case understanding rather than rejection of the construct. The same applies to the proposition that the QALY is dimensionally homogeneous. The knowledge base strongly supports the operational acceptance of QALYs while showing little recognition of the underlying requirements for ratio measurement and dimensional homogeneity.

The issue is not that ScHARR lacks technical sophistication. On the contrary, the knowledge base is likely to be highly sophisticated in economic modelling, evidence review, cost-effectiveness analysis, probabilistic sensitivity analysis, uncertainty assessment, and appraisal support. That is precisely why the finding matters. Sophistication in modelling does not compensate for failure in measurement. A model can be technically elaborate, internally consistent, and carefully validated against assumptions while still manipulating quantities that do not qualify as measures. If the inputs fail the requirements of representational measurement, the output inherits that failure.

This is the core problem for reference-case simulation models. They do not create measurement. They manipulate quantities already assumed to be valid. If utility values lack ratio properties, then

multiplying them by time to create QALYs is inadmissible. If QALYs lack dimensional homogeneity, then aggregating them across populations or using them as denominators in cost-effectiveness ratios cannot generate robust estimates of therapy impact. A simulation model cannot rescue a quantity that lacks measurement status. It can only extend the consequences of the original error across time, population subgroups, and hypothetical decision scenarios.

The ScHARR interrogation also reveals weak recognition of unidimensionality. The proposition that measures must be unidimensional receives only 0.15. This is significant because the utility instruments and quality-of-life measures commonly used in HTA are multidimensional. They combine mobility, pain, anxiety, self-care, usual activities and other domains into single values. Without demonstrating unidimensionality, there is no basis for claiming that these values represent measures of a single attribute. Aggregation may produce a number, but it does not produce a measure. The failure to recognize this distinction is one of the clearest signs of measurement inversion.

The treatment of latent attributes is equally problematic. HTA frequently incorporates outcomes such as pain, fatigue, physical functioning, mental health, treatment satisfaction, quality of life, and patient experience. These are latent attributes. They are not directly observable and require a measurement model to estimate possession of the attribute. Yet propositions concerning latent attribute measurement receive very low probabilities. The proposition that the Rasch logit ratio scale is the only basis for assessing therapy impact in latent traits receives only 0.05. The proposition that the outcome of interest for latent traits is possession of the trait receives only 0.20. The proposition that Rasch rules are identical to the axioms of representational measurement receives only 0.05.

These results indicate that Rasch measurement occupies almost no meaningful role in the ScHARR HTA knowledge base. This is not a minor omission. If patient-centered outcomes and quality-of-life claims are to be treated as quantitative, then a framework is required to transform ordinal observations into measures of latent attribute possession. Rasch measurement provides that framework. Without it, utility scores, composite indices, and patient-reported outcome totals remain numerical constructions. They may be useful descriptions or summaries, but they do not satisfy the requirements for quantitative measurement.

The absence of Rasch also explains why the distinction between manifest and latent attributes disappears. Manifest outcomes such as mortality, survival time, hospital admissions, adverse events, medication possession, and resource utilization can be directly observed and, where appropriately specified, assessed through linear ratio measures. Latent outcomes require Rasch logit ratio measures. These are different measurement problems. Yet the proposition that there are only two classes of measurement, linear ratio and Rasch logit ratio, receives only 0.05. The proposition that subjective responses can be transformed to measurement only through Rasch rules receives only 0.05. The knowledge base therefore appears to treat all numerical outputs as if they belong to a common analytic space, rather than recognizing that different attributes require different measurement frameworks.

This failure has direct implications for evidence review. ScHARR, like other UK appraisal support centers, is involved in the assessment of clinical evidence, economic models, and submitted claims.

Evidence review groups are expected to critique clinical evidence, mathematical models, the validity of results produced, and the interpretation of results. However, critique of model validity is not the same as critique of measurement status. A review group may examine model structure, parameter uncertainty, extrapolation assumptions, comparator selection, and sensitivity analyses. These are important tasks, but they do not answer the prior question: are the quantities entering the model measures?

This is where the ScHARR profile becomes especially important. The knowledge base appears highly capable of evaluating model plausibility, uncertainty and policy relevance, but far less capable of evaluating the measurement foundations of the quantities being modelled. The danger is that review becomes focused on assumptions within the reference case while the reference case itself escapes measurement scrutiny. The model is reviewed, but the measurement status of the utility, QALY, or latent outcome is assumed. Under these conditions, evidence review does not challenge measurement inversion; it institutionalizes it.

The propositions concerning false claims embedded in the reference-case framework receive strong endorsement. The knowledge base supports the treatment of time trade-off preferences, EQ-5D algorithms, QALYs, aggregated QALYs, reference-case simulations, and utility-based claims as legitimate elements of HTA practice. This pattern mirrors the wider NICE reference-case environment. The issue is not whether these constructs are familiar or routinely applied. They clearly are. The issue is whether the measurement conditions necessary to support them have been demonstrated. The interrogation indicates that those conditions are not recognized as governing constraints.

The result is a knowledge base in which arithmetic is normalized. Utility values are assigned, QALYs are constructed, ICERs are calculated, simulations are run, uncertainty is explored, and appraisal recommendations are supported. Each step appears methodologically rigorous when viewed from within the reference-case paradigm. Yet when judged against the standards of representational measurement, the sequence depends upon unproven assumptions. The framework assumes that numerical assignment is measurement. It assumes that arithmetic is admissible. It assumes that model outputs can be treated as evidence. These assumptions are precisely what measurement science requires to be demonstrated.

This is why the ScHARR interrogation should not be interpreted as merely identifying an educational gap. It identifies a foundational problem in the operational machinery of UK HTA. ScHARR is not peripheral to the system. It represents one of the institutions through which the reference case is implemented, reviewed, taught and reproduced. If the knowledge base gives little recognition to representational measurement, then the critique extends beyond academic theory. It reaches the processes through which technologies are appraised and reimbursement decisions are informed.

The profile also helps explain why the reference case has proven so resistant to challenge. When students, researchers, analysts and reviewers are trained within a framework that emphasizes modelling, evidence review and cost-effectiveness analysis but not measurement theory, the central assumptions of the paradigm become invisible. Utilities are not questioned because they are treated as standard inputs. QALYs are not questioned because they are treated as standard

outcomes. Simulation models are not questioned at the level of measurement because they are evaluated primarily for structure, uncertainty and plausibility. The omission of measurement science from the knowledge base therefore protects the paradigm from the most fundamental form of criticism.

The broader significance is clear. The ScHARR results are consistent with profiles observed for CHE, CRD, Oxford and Newcastle. Across UK HTA, the pattern is remarkably stable. Institutions differ in mission—economic evaluation, evidence synthesis, appraisal support, outcomes research, policy analysis—but the measurement profile remains the same. Representational measurement is absent. Ratio requirements are weakly endorsed. Rasch measurement is marginal. Latent attribute possession is neglected. QALYs, utilities and simulations remain central. This consistency points to systemic measurement inversion rather than local institutional weakness.

For claims of robust estimates of therapy impact, the implications are devastating. Robustness cannot be achieved by sensitivity analysis alone. It cannot be achieved by probabilistic modelling, scenario analysis, or alternative assumptions if the central quantities lack measurement status. A robust estimate requires a measure. If the input is a numerical construction rather than a measure, then the apparent robustness of the output is an artefact of the modelling framework. It is not evidence of therapeutic impact.

The interrogation therefore leads to a direct conclusion. ScHARR's HTA knowledge base demonstrates strong commitment to reference-case practice, evidence review and decision support, but weak recognition of the measurement science required to justify the framework it supports. This is measurement inversion in an operational setting. The reference case is not merely being taught; it is being normalized through appraisal practice. The result is a system in which the appearance of analytical rigor substitutes for the requirements of measurement.

For ScHARR, and for UK HTA more broadly, the challenge is therefore not to refine the reference case. It is to replace the assumptions of the reference case with a framework grounded in lawful measurement. Every claim must begin with an explicitly defined attribute. Manifest attributes must be assessed through linear ratio measures. Latent attributes must be assessed through Rasch logit ratio measures. Arithmetic must be constrained by demonstrated scale properties. Claims must be evaluable, replicable and falsifiable. Until these requirements are met, the knowledge base cannot support robust quantitative claims regarding therapy impact. The ScHARR interrogation therefore provides further evidence that the problem confronting UK HTA is not methodological weakness but paradigm failure.

ENDORSEMENT RESULTS: NEWCASTLE HTA GROUP

The interrogation of the Newcastle HTA knowledge base reveals a pattern that is now familiar across leading HTA institutions in the United Kingdom, Australia, Canada, and New Zealand. The pattern is not one of isolated methodological weakness or a simple omission in curriculum content. Rather, it reflects a deeper problem: the systematic displacement of measurement by arithmetic. Quantitative claims are accepted, manipulated, synthesized, and interpreted without first establishing whether the quantities involved satisfy the requirements of measurement. This is the defining characteristic of measurement inversion.

The Newcastle profile demonstrates strong endorsement of propositions consistent with the contemporary reference-case paradigm and weak endorsement of propositions grounded in representational measurement. The importance of this finding lies not simply in the individual probabilities but in the overall structure they reveal. The knowledge base demonstrates familiarity with evidence assessment, outcomes research, health technology assessment, and healthcare decision making. What remains largely absent are the principles that determine whether quantitative claims are scientifically legitimate.

The strongest endorsements occur where one would expect them. Time is correctly recognized as a ratio measure, receiving a probability of 0.95. Non-falsifiable claims should be rejected receives a probability of 0.70. The logit as the natural logarithm of the odds ratio receives moderate recognition. These findings indicate an awareness of empirical evaluation and scientific inquiry. Yet when attention shifts to the foundations of measurement, endorsement collapses.

The proposition that measurement precedes arithmetic receives a probability of only 0.10. Multiplication requires a ratio measure receives the same probability. The proposition that the axioms of representational measurement are required before arithmetic can be applied also receives only 0.10. These are not obscure technical propositions. They represent the foundations upon which all quantitative science depends. If measurement does not precede arithmetic, then there is no basis for determining whether arithmetic operations are admissible. Numbers may be manipulated, but there is no guarantee that those manipulations possess scientific meaning.

The implications are profound because contemporary HTA depends almost entirely upon arithmetic operations. Utility scores are averaged. Utility scores are multiplied by time to generate QALYs. QALYs are aggregated across populations. Costs are divided by QALYs to create cost-effectiveness ratios. Simulation models extend these calculations across hypothetical populations and time horizons. Every stage of the reference-case framework depends upon arithmetic. Yet the interrogation suggests little recognition that arithmetic itself requires prior measurement justification.

This omission becomes even more significant when considered alongside representational measurement. The proposition that arithmetic requires satisfaction of the axioms of representational measurement receives one of the lowest endorsement probabilities observed in the interrogation. This suggests that the relationship between empirical attributes, numerical representation, and admissible mathematical operations occupies little place within the Newcastle knowledge environment. The consequence is that quantitative claims are effectively treated as self-validating. Once a quantity is expressed numerically, it is assumed to support arithmetic operations without further inquiry into its measurement status.

The same pattern is evident in the treatment of unidimensionality. Measurement requires that an attribute represent a single dimension. Without unidimensionality there may be scoring, ranking, or numerical assignment, but there is no measurement. Yet the proposition that measures must be unidimensional receives only a probability of 0.15. This finding is important because many of the instruments routinely employed within HTA are multidimensional. Utility systems, quality-of-life instruments, and composite outcome measures combine multiple domains into a single numerical

output. Without explicit attention to unidimensionality there is no basis for determining whether these outputs represent measures or merely numerical constructions.

The low endorsement of latent attribute measurement is equally revealing. Contemporary HTA increasingly relies upon outcomes such as quality of life, pain, functioning, fatigue, psychological wellbeing, and patient satisfaction. These are not directly observable phenomena. They are latent attributes. Measurement of latent attributes requires a measurement model capable of estimating possession of the attribute. Yet the proposition that latent attributes require a measurement model receives only limited endorsement. Similarly, the proposition that the outcome of interest for latent attributes is possession of the attribute receives a low probability. These findings indicate little recognition that latent variables pose fundamentally different measurement challenges from observable outcomes.

The consequences extend directly to Rasch measurement. The propositions concerning Rasch logit ratio measurement receive among the lowest probabilities in the interrogation. The proposition that there are only two classes of measurement. linear ratio measurement for manifest attributes and Rasch logit ratio measurement for latent attributes, receives a probability of only 0.05. The proposition that transformation of subjective observations into measures requires Rasch rules receives the same probability. The proposition that the Rasch logit ratio scale provides the basis for assessing therapy impact in latent traits also receives only 0.05.

These findings are particularly important because they suggest that the conceptual framework required for latent measurement is almost entirely absent. Once Rasch measurement disappears from the knowledge base, latent outcomes are left without a scientifically defensible measurement framework. Utility instruments, patient-reported outcomes, quality-of-life measures, and preference-based assessments continue to be used, but the measurement problem itself effectively disappears from view.

The treatment of utilities and QALYs illustrates this phenomenon clearly. The interrogation indicates strong rejection of propositions challenging the ratio properties of QALYs and utility-based constructs. The implication is that the knowledge base continues to treat utilities and QALYs as though their measurement properties were already established. Yet the propositions concerning ratio measurement, representational measurement, and admissible arithmetic receive consistently low endorsement. The contradiction is obvious. The outputs of the reference-case framework are accepted while the conditions necessary to justify those outputs are neglected.

This contradiction lies at the heart of measurement inversion. Measurement inversion occurs when arithmetic is accepted before measurement is established. Instead of asking whether a quantity satisfies the requirements necessary to support a particular mathematical operation, the operation is performed first and measurement is assumed to follow automatically. The distinction between a number and a measure disappears. Numerical construction is mistaken for measurement.

The significance of this finding is especially important in an institution whose mission includes evidence assessment and policy evaluation. The assumption underlying evidence-based decision making is that evidence consists of valid observations and measures. Yet the interrogation raises a prior question: what is the measurement status of the quantities being evaluated? Evidence

synthesis can summarize findings across studies. Meta-analysis can combine results from multiple investigations. Technology assessment can compare interventions across populations and settings. None of these activities can create measurement where measurement does not exist.

This point cannot be overstated. A systematic review of non-measures does not transform non-measures into measures. A meta-analysis of utility scores does not establish ratio properties. Aggregation of QALYs does not create dimensional homogeneity. Evidence synthesis can strengthen confidence in a finding only if the quantity being synthesized already satisfies the requirements of measurement. Otherwise, the review merely synthesizes the consequences of measurement failure.

This observation has important implications for Newcastle's role within the HTA environment. The institution contributes to evidence reviews, policy analysis, health services research, and healthcare decision making. Yet the interrogation suggests that the concepts necessary to evaluate the measurement status of the evidence being reviewed occupy only a marginal position within the knowledge base. The result is a framework that may be highly sophisticated in its treatment of evidence while remaining largely silent regarding the measurement foundations of that evidence.

The findings also provide insight into the persistence of the reference-case paradigm. It is often assumed that the longevity of utilities, QALYs, and simulation models demonstrates their scientific legitimacy. The interrogation suggests a different explanation. The concepts necessary to challenge the paradigm are themselves largely absent from the knowledge base. If researchers are not exposed to representational measurement, ratio measurement, latent attribute measurement, unidimensionality, and Rasch measurement, they cannot be expected to evaluate the assumptions underlying utilities and QALYs. The paradigm survives not because its foundations have been demonstrated but because the intellectual tools required to evaluate those foundations remain largely absent.

This conclusion is reinforced by comparison with other leading HTA institutions. Similar interrogation profiles have been observed at CHE, CRD, Oxford, ScHARR, and across major centers in Australia, Canada, and New Zealand. The same concepts receive low endorsement. The same assumptions receive high endorsement. The same absence of measurement science recurs regardless of institutional mission. The consistency of these findings suggests that the problem is not local but systemic. Newcastle does not represent an exception. It represents another manifestation of a broader HTA culture in which arithmetic has displaced measurement as the starting point for quantitative claims.

The implications for the scientific status of the reference-case paradigm are substantial. The reference case presents itself as a framework capable of generating robust estimates of therapeutic value. Yet the interrogation suggests that the concepts required to establish the measurement properties of its central constructs are largely absent. Utility scores are treated as though they possess ratio properties. QALYs are treated as though they are dimensionally homogeneous measures. Simulation models are treated as though they generate evidence. However, the measurement foundations required to support these claims receive little endorsement.

This is not merely an interesting methodological observation. It strikes directly at the scientific legitimacy of the framework. A paradigm that claims to generate quantitative estimates of therapy impact must be able to demonstrate that the quantities entering those estimates satisfy the requirements of measurement. If it cannot do so, then the resulting claims lose their scientific standing. The issue is no longer one of methodological refinement or improved modelling techniques. The issue is whether the framework satisfies the minimum standards required for quantitative science.

The Newcastle interrogation therefore points to a clear conclusion. The knowledge base demonstrates strong engagement with evidence, evaluation, and policy analysis but limited engagement with the principles of measurement science. The resulting pattern is one of measurement inversion. Quantitative claims are accepted while the requirements necessary to justify those claims remain largely absent. In doing so, the knowledge base reproduces the central weakness of the contemporary HTA paradigm itself. The problem is not simply that measurement science has been neglected. The problem is that the neglect has become institutionalized. Once this occurs, the resulting framework can no longer be regarded as a secure foundation for claims regarding therapeutic value. The interrogation therefore provides further evidence that the challenge facing HTA is not one of reform but of reconstruction.

OVERALL: MEASUREMENT INVERSION DOMINATES

This is indeed the strongest result so far, and arguably stronger than the corresponding analyses for Australia, New Zealand, and Canada because of the extraordinary uniformity across five institutions that collectively represent the intellectual core of UK HTA. The issue is not simply that endorsement of measurement principles is low. The issue is that the same pattern appears repeatedly across institutions whose missions differ substantially. CHE focuses on economic evaluation, CRD on evidence synthesis, Oxford on outcomes research and health economics, ScHARR on technology appraisal, and Newcastle on applied HTA and policy analysis. Yet all five arrive at essentially the same profile.

The first feature of the results is the near-total absence of representational measurement. The proposition that meeting the axioms of representational measurement is required for arithmetic receives a probability of only 0.10 in every center. Equally striking, the proposition that measurement precedes arithmetic also receives 0.10 in every center. This level of consistency is remarkable. These are not controversial propositions. They represent the foundational requirements of quantitative science. Yet across the leading HTA centers in the United Kingdom they occupy virtually no visible place within the knowledge base.

The same pattern is evident for ratio measurement. Multiplication requires a ratio measure receives a probability of only 0.10 in every institution. At the same time, propositions central to the utility-QALY framework receive endorsement probabilities between 0.85 and 0.95. The contradiction is obvious. The arithmetic operations required for utility construction, QALY generation, aggregation, and cost-effectiveness analysis are accepted, while the measurement requirements necessary to justify those operations are largely absent. Arithmetic is endorsed while measurement is neglected.

Unidimensionality presents an equally striking result. The proposition that measures must be unidimensional receives a probability of only 0.15 across all five centers. Yet unidimensionality is a prerequisite for measurement. Without it, there may be scoring, ranking, aggregation, or indexing, but there is no measurement. The uniformity of this result suggests that the distinction between multidimensional descriptive systems and genuine measures is largely absent from the intellectual environment supporting UK HTA.

Perhaps the most revealing findings concern latent attribute measurement. Across all five institutions, the proposition that there are only two classes of measurement, linear ratio measurement for manifest attributes and Rasch logit ratio measurement for latent attributes, receives a probability of only 0.05. The proposition that transforming subjective responses into measurement is only possible through Rasch rules also receives 0.05. The proposition that the Rasch logit ratio scale provides the basis for assessing therapy impact in latent traits receives 0.05 across every institution. The proposition that Rasch measurement embodies the axioms of representational measurement likewise receives 0.05.

These are among the lowest probabilities observed in any national assessment undertaken to date. They indicate not merely a lack of familiarity with Rasch measurement but the virtual absence of the conceptual framework required to measure latent attributes. This finding is particularly important because contemporary HTA increasingly depends upon latent outcomes: quality of life, pain, fatigue, functioning, psychological wellbeing, treatment satisfaction, and patient experience. If the measurement framework required for these outcomes is absent, then the scientific basis for claims regarding their quantitative assessment becomes highly questionable.

The treatment of the manifest-latent distinction reinforces this conclusion. The proposition that the outcome of interest for latent traits is possession of the trait receives probabilities between 0.20 and 0.25. Recognition that manifest and latent attributes require different forms of ratio measurement is effectively absent. This is a crucial finding because it identifies one of the central conceptual failures within contemporary HTA. Manifest outcomes such as survival time, hospital admissions, adverse events, treatment persistence, and resource utilization can support direct observation and linear ratio measurement. Latent outcomes require Rasch logit ratio measurement. Once this distinction disappears, all numerical outputs begin to appear interchangeable. The consequence is that ordinal responses, utility scores, latent constructs, and observable outcomes are routinely treated as though they occupy the same measurement space.

The findings for falsification provide an interesting contrast. Non-falsifiable claims should be rejected receives probabilities ranging from 0.65 to 0.80. Similarly, the logit as the natural logarithm of the odds ratio receives moderate endorsement. These results suggest that the knowledge bases retain some appreciation of empirical inquiry and scientific testing. Yet this only deepens the contradiction. Falsifiability is recognized while the measurement conditions necessary to support falsifiable quantitative claims are neglected. The result is a framework that values evidence but gives little attention to the measurement status of the evidence itself.

Taken together, the findings provide perhaps the clearest evidence yet of measurement inversion. Across all five centers, true propositions grounded in measurement science receive consistently low endorsement, while false propositions embedded within the reference-case paradigm receive

consistently high endorsement. This is not random omission. It is systematic reversal. The knowledge bases repeatedly favor propositions that conflict with representational measurement while failing to endorse propositions that reflect its principles.

The significance of this consistency should not be underestimated. These five centers collectively represent the intellectual foundation of UK HTA. They have contributed to the development of NICE methods, evidence synthesis, economic evaluation, technology appraisal, outcomes research, and health policy. If the concepts required to establish lawful measurement are absent here, then the problem extends to the foundations of the reference-case paradigm itself.

This is why the results are so important. They provide an explanation for the persistence of utilities, QALYs, cost-effectiveness ratios, and simulation models despite their failure to satisfy the requirements of measurement. The concepts necessary to identify the defect are themselves largely absent from the educational and research environment. Curriculum inversion reproduces measurement inversion. Researchers are trained to construct and interpret quantitative outputs without first being trained to determine whether the quantities entering those analyses are measures.

The conclusion is unavoidable. The UK HTA knowledge base does not merely contain isolated examples of measurement failure. It demonstrates a systematic and reproducible pattern of measurement inversion. The concepts required to establish valid quantitative claims occupy a marginal position, while the assumptions underpinning the reference-case paradigm dominate the analytical framework. For a field that claims to generate robust estimates of therapeutic value and therapy impact, the implications are devastating. The issue is no longer one of methodological refinement. The issue is paradigm failure. The intellectual foundations of the reference case have been shown to rest upon assumptions regarding measurement rather than measurement itself.

CONCLUSION: THE FUTURE FOR HTA

The findings reported in this assessment leave little room for ambiguity. The contemporary framework of utilities, QALYs, cost-effectiveness ratios, and reference-case simulation models has no scientific future. This conclusion does not arise from disagreement over methodology, competing analytical preferences, or calls for further refinement. It follows directly from the evidence presented in the interrogation studies. Across the leading UK research centers there is little recognition of the standards governing measurement, the properties of measurement scales, the requirements of ratio measurement, the distinction between manifest and latent attributes, or the axioms of representational measurement upon which all quantitative claims depend.

This absence of measurement science is not a minor educational deficiency. It exposes a far more serious problem. The architects of contemporary HTA, together with their academic successors, constructed and defended a framework whose central analytical operations require measurement properties that were never demonstrated. Utilities were treated as though they possessed ratio properties. QALYs were created through multiplication without first establishing that multiplication was admissible. Cost-effectiveness ratios were generated from quantities whose measurement status remained unresolved. Reference-case simulation models then extended these

assumptions across hypothetical populations and lifetime horizons. The result is a framework built upon assumptions regarding measurement rather than measurement itself.

Under these circumstances, the conclusion is unavoidable: the reference-case paradigm represents a case of paradigm failure. The issue is not that the framework can be repaired through improved utility instruments, revised modelling standards, greater computational sophistication, or more elaborate evidence synthesis. The defect lies deeper. The framework was established without satisfying the conditions necessary for quantitative science. It did not begin with measurement and later lose its way. It began without measurement. Four decades of methodological elaboration have merely extended and institutionalized the original error.

For this reason, the future of HTA cannot be found within the reference-case paradigm. There is no defensible path forward that preserves utilities, QALYs, and simulation-based cost-effectiveness claims. Once the requirements of representational measurement are restored, these constructs collapse. They cannot be rescued by assumption, convention, consensus, or administrative necessity. The laws governing measurement do not yield to professional preference.

What emerges in place of the reference case is not the abandonment of HTA but its reconstruction. The essential elements of a scientific framework are already clear. Every assessment begins with an explicitly defined attribute. That attribute must then be classified as either manifest or latent. Manifest attributes support direct observation and require linear ratio measurement. Latent attributes require a measurement model capable of estimating possession of the attribute and therefore require Rasch logit ratio measurement. These are the only two forms of measurement capable of supporting quantitative claims regarding therapy impact.

Once measurement is established, arithmetic becomes lawful rather than assumed. Claims can be expressed in terms of measurable change in attributes. Evidence becomes evaluable. Competing claims can be replicated. Most importantly, all claims become capable of falsification. This restores HTA to the standards that govern normal science. Measurement precedes arithmetic. Quantitative claims are grounded in measures rather than numerical constructions. Evidence is generated through observation rather than simulation. Theories and claims survive only so long as they withstand empirical testing.

The future of HTA therefore lies not in defending the status quo but in replacing it. The transition is from a framework dominated by utilities, QALYs, and simulation models to one grounded in attributes, ratio measurement, and falsifiable claims. The choice is stark. One path preserves a paradigm that has failed the standards of measurement science. The other establishes a discipline capable of producing credible and scientifically defensible claims regarding therapeutic impact. The evidence presented here leaves little doubt as to which path should be chosen.

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

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

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

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