MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: CLOSED EPISTEMIC REPRODUCTION - HOW ACADEMIC HTA KNOWLEDGE BASES IN THE UNITED STATES SUSTAIN MEASUREMENT INVERSION

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

LOGIT WORKING PAPER No 757 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

This study examines the structure of seven health technology assessment (HTA) knowledge bases within US academic research centers through the lens of the philosophy of science and representational measurement. Using a standardized 24-item canonical diagnostic, five university-based HTA knowledge domains were interrogated to assess whether they recognize the axioms required for valid measurement and the conditions necessary for falsification. Each statement was evaluated within the context of the knowledge base, with responses expressed as categorical probabilities and transformed into normalized logits to enable comparison across domains.

The results demonstrate a consistent and reproducible pattern. Statements that are true under the axioms of measurement, such as the requirement for unidimensionality, dimensional homogeneity, and the precedence of measurement over arithmetic are weakly endorsed, with probabilities typically between 0.10 and 0.25 and corresponding negative logits. In contrast, statements that are false under these axioms but embedded within HTA practice such as the treatment of QALYs as ratio measures and the aggregation of utility-based constructs are strongly endorsed, with probabilities between 0.80 and 0.90 and positive logits. This pattern mirrors findings from Canadian HTA knowledge bases, indicating that the observed structure is not jurisdiction-specific but reproducible across independent domains.

A distinctive feature of the US results is the partial endorsement of scientific principles such as falsifiability, which does not extend to the core constructs of HTA. This disconnect supports the interpretation of HTA as a system of closed epistemic reproduction, in which methods and assumptions are transmitted and reinforced without engagement with the conditions required for their validation.

The study concludes that HTA, as currently practiced in these seven research centers, operates outside the framework of scientific measurement. Its outputs are numerical but not measurable, and its core constructs are not subject to falsification. Addressing this requires a transition to measurement-based evaluation grounded in the axioms of representational measurement and the principles of scientific inquiry.

INTRODUCTION – MEASUREMENT, SCIENCE, AND THE PERSISTENCE OF HTA PRACTICE

Health technology assessment (HTA) presents itself as a quantitative science. Its claims are expressed in numerical form, its methods rely on formal models, and its outputs are used to inform decisions regarding the allocation of healthcare resources. This appearance of scientific rigor rests on an implicit assumption: that the quantities manipulated within HTA such as utilities, QALYs, and cost-effectiveness ratios possess the properties required for measurement. Without this assumption, the use of arithmetic loses its meaning. Numbers may be generated, but they do not represent measurable attributes.

The philosophy of science provides a clear framework for evaluating such claims. Since the scientific revolution, the legitimacy of quantitative inquiry has depended on two conditions: the establishment of measurement and the possibility of falsification. Measurement ensures that attributes are represented on scales with defined properties, allowing for meaningful comparison and transformation. Falsification ensures that claims can be subjected to empirical test and, if necessary, rejected. Together, these conditions support the evolution of objective knowledge. Where they are absent, numerical reasoning may persist, but it does so outside the boundaries of science.

The present study examines HTA within this framework. Using a standardized 24-item canonical diagnostic grounded in the axioms of representational measurement, a series of knowledge bases associated with Australian academic HTA research centers have been interrogated. Each knowledge base is treated as a defined domain, encompassing its published outputs, methodological conventions, and implicit assumptions. The interrogation evaluates whether propositions that are true under the axioms of measurement are recognized, and whether propositions that violate those axioms are nevertheless endorsed. Responses are expressed as categorical probabilities and transformed into normalized logits, allowing for comparison across statements and across domains.

The purpose of this paper is not only to present the results of these interrogations, but to address a deeper question: why does a consistent pattern emerge in which the conditions required for measurement are not recognized, while incompatible constructs are reinforced? The findings demonstrate that this pattern is not incidental. It is stable across institutions and reproducible across jurisdictions. This raises a philosophical problem. If HTA operates with numerical constructs that do not meet the requirements of measurement, and if these constructs are not subjected to falsification, then on what basis does the framework persist?

The argument developed in this paper is that HTA represents a case of closed epistemic reproduction: a system in which methods and constructs are transmitted, reinforced, and repeated without engagement with the conditions required for their validation. Understanding this persistence requires moving beyond technical critique to a consideration of how knowledge systems are sustained, how they resist challenge, and how they can be replaced.

FROM INDIVIDUAL KNOWLEDGE BASE INTERROGATION TO CONSENSUS

The starting point for this analysis is a series of individual interrogations of Australian HTA academic knowledge bases. Each interrogation was conducted independently, applying a standardized 24-item canonical diagnostic grounded in the axioms of representational measurement. The purpose of these interrogations was to determine, within each domain, whether the conditions required for valid measurement were recognized and whether propositions incompatible with those conditions were nevertheless endorsed. At this stage, no assumption was made regarding the relationship between knowledge bases. Each was treated as a separate and self-contained domain.

The results of these individual interrogations, taken in isolation, were already informative. In each case, statements that are true under the axioms of measurement—such as the requirement for unidimensionality, the necessity of dimensional homogeneity, and the principle that measurement must precede arithmetic—were weakly endorsed. Conversely, statements that are false under those same axioms, but embedded within HTA practice—such as the treatment of QALYs as ratio measures and the aggregation of utility-based constructs—were strongly endorsed. This pattern was observed consistently within each knowledge base. However, at the level of individual analysis, the interpretation remained open. It was possible to attribute the results to local factors: institutional emphasis, methodological preferences, or variations in training.

The transition to synthesis removes this ambiguity. When the results from multiple knowledge bases are brought together and examined in terms of probability and logit ranges, a different picture emerges. The pattern observed within each domain is reproduced across domains, with minimal variation in both magnitude and direction. Statements that are true under measurement theory occupy a consistently negative range of logits; statements that are false occupy a consistently positive range. There is no instance in which a knowledge base reverses this pattern. The convergence is not approximate; it is structural.

This convergence justifies the introduction of the concept of consensus. It is important to distinguish the meaning of consensus in this context from its conventional use. This is not consensus in the sense of agreement among researchers or institutions. There is no evidence of coordinated positions or explicit alignment. Rather, consensus refers to the uniformity of knowledge base structure when subjected to a common interrogation. Independent domains, operating under different institutional arrangements and pursuing different research agendas, produce the same pattern of endorsement across the same set of statements. The agreement is implicit, not declared.

The strength of this consensus lies in its reproducibility. The use of a fixed diagnostic instrument, constrained probability categories, and a standardized transformation to normalized logits ensures that comparisons across knowledge bases are valid. Differences in results can therefore be attributed to differences in the knowledge base, not to variation in method. When such differences fail to appear—when the same pattern is observed repeatedly—the conclusion is that the underlying structure is shared.

This shared structure is not limited to Australia. The same pattern has been documented in Canadian HTA knowledge bases, where a similar set of interrogations produced comparable ranges of probabilities and logits. Despite differences in policy environments, institutional configurations, and academic traditions, the direction and magnitude of endorsement remain aligned. This cross-jurisdictional consistency strengthens the interpretation of consensus. It indicates that the pattern is not contingent on local conditions, but reflects a more general property of HTA as a field.

At this point, the focus of the analysis shifts. The question is no longer whether individual knowledge bases recognize the conditions required for measurement. That question has been answered in the negative. The question is whether the absence of recognition is variable or systematic. The results show that it is systematic. The same propositions are weakly endorsed across all domains; the same incompatible constructs are strongly reinforced. The pattern is stable, reproducible, and resistant to variation.

This raises a further question. If independent knowledge bases, operating in different contexts, produce the same pattern, then the explanation cannot lie in isolated methodological choices. It must lie in the way the knowledge base itself is constituted and transmitted. The concept of consensus therefore serves as a bridge. It marks the point at which individual observations coalesce into a structural finding, and it prepares the ground for the central problem addressed in this paper: not only that measurement is absent, but that its absence is consistently reproduced.

The remainder of the analysis builds on this foundation. Having established that the pattern is one of consensus rather than variation, the focus turns to its implications and, more importantly, to its persistence. The key issue is not simply that HTA fails to meet the conditions required for measurement, but that it does so in a way that is stable across domains and resistant to correction. Understanding this persistence requires moving beyond description to explanation.

THE US HTA RESEARCH LANDSCAPE

The United States health technology assessment (HTA) research landscape is characterized by a diverse but tightly interconnected set of academic centers that play a central role in the development, application, and dissemination of economic evaluation and outcomes research methods. These centers are typically embedded within major research universities, often located in schools of public health, medicine, or policy, and are closely aligned with healthcare decision-making in both public and private sectors. Their outputs include cost-effectiveness analyses, decision-analytic models, real-world evidence studies, and policy-relevant evaluations. While differing in institutional mission and research emphasis, they collectively define the academic knowledge base from which HTA-related practice in the United States is drawn.

Among the most prominent of these are the USC Schaeffer Center for Health Policy and Economics, the Harvard T.H. Chan School of Public Health Center for Health Decision Science, the University of Washington CHOICE Institute, the Duke-Margolis Institute for Health Policy, the University of Pennsylvania Leonard Davis Institute of Health Economics, the Stanford Health Policy Center, and the Health Economics Group at the Johns Hopkins Bloomberg School of Public Health. These centers vary in their specific focus—some emphasizing methodological development in decision science, others concentrating on policy engagement, real-world evidence, or global health—but all operate within a common analytical framework centered on cost-effectiveness analysis and the use of preference-based outcome measures.

A defining feature of the U.S. HTA landscape is its diffuse but powerful alignment with payer, regulatory, and policy environments. Unlike jurisdictions with a single national decision-making body, the United States operates through a combination of federal agencies, private payers, and advisory organizations. Despite this fragmentation, there is a strong convergence around the use of economic evaluation, particularly cost per QALY frameworks, as a reference point for assessing value. Academic centers align their methods with these expectations, ensuring that their outputs are relevant to stakeholders involved in pricing, coverage, and reimbursement decisions. As a result, there is an implicit standardization across institutions, with utilities, QALYs, and cost-effectiveness ratios serving as the common currency of evaluation.

This alignment has important implications for the structure of the knowledge base. The use of shared methodological conventions, modeling approaches, and outcome measures promotes consistency across institutions. Researchers are trained within programs that emphasize accepted HTA techniques, and their work is disseminated through journals and policy channels that reinforce adherence to these methods. Over time, this creates a stable and internally coherent framework in which variation occurs at the level of application—different diseases, interventions, and datasets—but not at the level of underlying constructs.

Despite this diversity in application, the core methodological elements remain constant. Economic evaluations rely on preference-based measures of health-related quality of life, typically derived from multiattribute instruments, and these measures are combined with time-based outcomes to generate QALYs. Decision-analytic models are used to extrapolate beyond observed data, incorporating assumptions regarding disease progression, treatment effects, and patient behavior. Increasingly, these models are supplemented by real-world evidence drawn from large-scale datasets, including claims and electronic health records. The outputs are expressed in numerical form and interpreted as evidence to support decision-making.

The knowledge base that emerges from this landscape is therefore both extensive and coherent. It encompasses a wide range of empirical studies, methodological developments, and policy analyses, all operating within a shared conceptual framework. This framework is rarely articulated in terms of its measurement properties, but it is implicitly defined by the requirements of HTA practice and the expectations of decision-makers. The interrogation of this knowledge base, using the canonical diagnostic, does not focus on the breadth of output, but on the underlying assumptions that govern the use of numbers within this framework.

The inclusion of multiple U.S. academic centers in this analysis serves to test whether these assumptions vary across institutional contexts. Given the diversity of the U.S. system—spanning policy-oriented institutes, methodological centers, and global health research groups—one might expect variation in how measurement principles are recognized or applied. However, as the results demonstrate, this variation does not occur at the level of measurement. The knowledge base, while diverse in its applications, exhibits a consistent structure in its treatment of the conditions required for measurement.

This consistency is central to the argument that follows. The U.S. HTA research landscape provides a robust and heterogeneous setting in which to examine whether the absence of measurement is an isolated phenomenon or a systemic property. By selecting knowledge bases that differ in institutional form and analytical emphasis, the analysis distinguishes between local variation and structural uniformity. The findings indicate that it is the latter that prevails.

·INTERROGATION FRAMEWORK AND ANALYTICAL METHOD

The analysis presented in this paper is based on a standardized interrogation framework designed to evaluate whether a knowledge base recognizes the conditions required for valid measurement. The framework is grounded in the axioms of representational measurement and operationalized through a fixed diagnostic instrument comprising 24 canonical statements. Each statement is constructed such that it is unequivocally true or false under these axioms. The purpose is not to solicit opinion or interpretation, but to determine whether the knowledge base, as represented in its corpus of outputs and methodological conventions, reinforces or rejects propositions that are fundamental to measurement.

Each US HTA research center is treated as a defined knowledge domain. This domain includes its published research, methodological approaches, teaching materials, and the implicit assumptions that can be inferred from these sources. The interrogation is conducted by presenting the canonical statements sequentially, with the knowledge base instructed to evaluate each statement in the context of its own established practices. The constraint is important: the knowledge base is not invited to reinterpret the statement or introduce external qualifications, but to assess it as it would be understood within the domain. This ensures that the response reflects the internal structure of the knowledge base rather than the flexibility of language.

Responses are expressed as categorical probabilities drawn from a fixed and deliberately limited set. These categories—such as 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.55, 0.60, 0.65, 0.75, 0.80, 0.85, 0.90, and 0.95—are selected to avoid spurious precision while allowing meaningful differentiation in the strength of endorsement. The assigned probability represents the extent to which the knowledge base supports the proposition embodied in the statement. Low probabilities indicate weak endorsement or effective rejection; high probabilities indicate strong endorsement. Intermediate values capture limited or partial recognition.

These categorical probabilities are then transformed into normalized logits using the standard log-odds transformation. The logit, defined as the natural logarithm of the odds ratio, maps bounded probability values onto a continuous scale. For the purposes of this analysis, logits are normalized to a fixed range (±2.50) using established rounding conventions. This normalization ensures comparability across interrogations and prevents extreme values from distorting interpretation. Equal differences in logits correspond to equal differences in the strength of endorsement, allowing the identification of patterns that are not apparent from ordinal probability categories alone.

Within a given interrogation, the set of probability–logit pairs across the 24 statements constitutes a profile of the knowledge base. Differences in these values indicate the relative strength with which individual propositions are supported or rejected. More importantly, the distribution of logits across true and false statements reveals whether the knowledge base aligns with or diverges from the axioms of measurement. A profile in which true statements are strongly endorsed and false statements rejected would indicate alignment with measurement principles. The inverse pattern indicates divergence.

Across interrogations, the use of a fixed diagnostic instrument, constrained probability categories, and a consistent logit transformation ensures that comparisons are valid. Because each knowledge base is evaluated under identical conditions, differences in results can be attributed to differences in the knowledge base rather than variation in method. When multiple knowledge bases produce similar profiles, the interpretation is that they share an underlying structure.

It is important to emphasize that the objective of this framework is not to produce a summary score, ranking, or index. The focus is on structure, not aggregation. Each statement is evaluated independently, and the resulting probabilities and logits are interpreted in relation to the truth value of the statement under representational measurement. The absence of a composite score avoids the introduction of additional assumptions regarding weighting or dimensionality, which would themselves require justification.

The framework also incorporates an implicit test of falsifiability. Statements concerning the rejection of non-falsifiable claims and the role of models in generating testable propositions are included alongside measurement axioms. This allows the interrogation to assess not only whether measurement is recognized, but whether the knowledge base acknowledges the conditions required for scientific testing. The relationship between these elements is central to the interpretation of results.

The application of this framework to Australian HTA research centers follows the same protocol as that used in the Canadian analysis. This ensures that the results are directly comparable across jurisdictions. Any observed similarities or differences can therefore be interpreted as properties of the knowledge base rather than artifacts of the method. The consistency of the approach is essential for identifying whether the patterns observed are local or structural.

In summary, the interrogation framework provides a systematic and reproducible means of evaluating the alignment of HTA knowledge bases with the axioms of representational measurement. By combining categorical probabilities with normalized logits, it allows for both qualitative and quantitative assessment of endorsement patterns. When applied across multiple domains, it reveals whether those patterns are variable or consistent, and thereby supports the transition from individual analysis to structural interpretation.

RESULTS: A CONSESUS PATTERN OF MEASUREMENT INVERSION

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

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

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY RANGE	NORMALIZED LOGIT (IN RANGE +/- 2.50) RANGE
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40

MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.60
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.80 to 0.85	+1.40 to +1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.80 to 0.85	+1.40 to +1.75
THE QALY IS A RATIO MEASURE	0	0.85 to 0.90	+1.75 to +2.20
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.85	+1.75
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85 to 0.90	+1.75 to +2.20
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15 to 0.20	-1.40 0 -1.75
QALYS CAN BE AGGREGATED	0	0.85 to 0.90	+1.75 to +2.20
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.60 to 0.65	+0.70 to +0.85
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85 to 0.90	+1.75 to +2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.60 to 0.65	+0.70 to +0.85
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50

A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.25 to 0.35	-1.90 to -1.25
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.90
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

Note: Where no range is reported all six knowledge bases were allocated to the same categorical endorsement probability.

The results presented in Table 1 provide a consolidated interrogation of seven leading U.S. academic centers engaged in health technology assessment (HTA), economic evaluation, and policy analysis. These include institutions with significant influence over methodological development, training, and policy guidance. The expectation, if HTA were evolving as a scientific discipline, would be that variation in institutional focus, disciplinary background, and policy engagement would produce diversity in the recognition and application of measurement principles. The results demonstrate the opposite. What emerges is not diversity, but remarkable uniformity.

This uniformity is not limited to general tendencies. It is expressed with numerical precision. For each of the 24 canonical statements, the probability ranges are narrow, often collapsing to a single value or a very tight interval. The corresponding logits exhibit the same stability. This is not a set of loosely aligned perspectives. It is **a** coherent and reproducible structure of belief**.**

The most striking feature of the table is the systematic separation between statements that are true under the axioms of representational measurement and those that are false. The former are consistently assigned low probabilities and negative logits. The latter are consistently assigned high probabilities and positive logits. This is not random variation. It is directional measurement inversion.

Consider the foundational axioms. The requirement that measurement must precede arithmetic is assigned a probability of 0.10, with logit of -2.20. The requirement that multiplication requires a ratio scale is fixed at 0.10 (-2.20). The recognition that measurement must be unidimensional is fixed at 0.15 (-1.60). These are not marginal deviations. They represent systematic rejection of the conditions required for measurement.

At the same time, statements that violate these axioms are strongly endorsed. The QALY is treated as a ratio measure with probabilities of 0.85 and 0.90. Its dimensional homogeneity is similarly endorsed. The aggregation of QALYs is accepted. Preference-based algorithms are assumed to generate interval scales. These endorsements are not tentative. They occupy the upper end of the probability scale, with logits approaching +2.20.

The contradiction is immediate and irreconcilable. The system rejects the requirement that multiplication be restricted to ratio scales while simultaneously endorsing the multiplication of utility values by time. It rejects the necessity of measurement preceding arithmetic while engaging in arithmetic operations on constructs that lack measurement properties. The contradiction is not recognized within the knowledge base. It is embedded within it.

The absence of variation across centers strengthens the interpretation. One might expect that institutions with a stronger policy orientation, such as those engaged in regulatory science or real-world evidence, would exhibit greater sensitivity to measurement requirements. Alternatively, centers with a stronger methodological focus might be expected to recognize the limitations of utility-based constructs. Neither expectation is supported. The ranges are narrow. The pattern is invariant.

This invariance is particularly evident in the treatment of Rasch measurement. Statements identifying Rasch as the only valid method for transforming ordinal responses to interval measurement are assigned probabilities of 0.05 across all centers. The corresponding logits are fixed at -2.50. This represents complete non-recognition. There is no variation, no partial endorsement, no evidence of engagement. The same applies to the identification of two classes of measurement: linear ratio and Rasch logit ratio. Again, probabilities of 0.05 and logits of -2.50. The implication is clear: the knowledge base does not incorporate a framework for the measurement of latent constructs.

At the same time, the system exhibits selective recognition of scientific principles. The rejection of non-falsifiable claims is moderately endorsed, with probabilities between 0.60 and 0.65. The definition of the logit is also recognized. These responses indicate that the language of science is present. Concepts such as falsifiability and mathematical definition are acknowledged. However, this recognition is compartmentalized. It is not applied to the core constructs of HTA.

This compartmentalization is critical. It allows the system to present itself as scientifically grounded while avoiding engagement with the implications of measurement theory. The presence of statistical rigor, large datasets, and complex models reinforces this perception. Yet none of these addresses the prior requirement of measurement. Without measurement, the outputs of these evaluations remain numerical representations without measurable attributes. It is just numerical storytelling,

The stability of the ranges across all seven centers confirms that this is not a problem of individual misunderstanding or isolated error. It is a structural property of the HTA knowledge base. The consistency of the probability–logit mapping indicates that the same conceptual framework is being reproduced across institutions. This reproduction is not accidental. It reflects shared training, shared publication outlets, shared policy frameworks, and shared assumptions about what constitutes valid evidence.

IMPLICATIONS: CLOSED EPISTEMIC REPRODUCTION AND THE FAILURE OF MEASUREMENT

The US results, taken together with the Australian and Canadian analysis, point to a conclusion that extends beyond individual jurisdictions. What is observed is not simply a recurring pattern, but a system that reproduces itself in the absence of measurement. The consistency of probabilities and logits across independent knowledge bases, combined with the persistence of the same directional structure demonstrates that HTA operates within a closed epistemic framework. The implications of this are not technical. They are structural, scientific, and policy-relevant.

The first implication is that HTA, as currently practiced, does not meet the conditions required for scientific measurement. This is not an abstract concern. The results show that statements defining those conditions such as unidimensionality, dimensional homogeneity, and the requirement that measurement precede arithmetic are weakly endorsed across all domains. This establishes a clear disconnect: arithmetic is applied where measurement has not been established. The consequence is that the outputs of HTA cannot be interpreted as measures in any scientific sense.

The second implication is that the framework is self-reinforcing. The US results show that even where there is partial recognition of scientific principle, most notably in the endorsement of falsifiability this recognition does not extend to the core constructs of HTA. QALYs, utilities, and cost-effectiveness ratios continue to be treated as meaningful quantities, despite their failure to meet the requirements of representational measurement. This indicates that the framework is not only closed, but selectively closed. It admits the language of science, but excludes its application where it would challenge established practice. The result is a system in which critique is acknowledged in principle but neutralized in practice.

This leads to a third implication: the absence of falsification. In a scientific system, claims are subject to test. Measurement provides the basis for this process, allowing propositions to be evaluated against observed reality. In HTA, the core constructs are not falsified because they are not measurable. QALY-based claims and cost-effectiveness ratios are outputs of models that incorporate assumptions about preferences, utilities, and disease progression. These assumptions are not tested as measures; they are embedded within the model. As a result, the system evolves through elaboration rather than through refutation. The interrogation results confirm this: the rejection of non-falsifiable claims is weakly to moderately endorsed, but the constructs themselves remain untouched.

A fourth implication concerns the interpretation of evidence. HTA presents its outputs as quantitative evidence to inform decision-making. The use of numbers such as costs, QALYs, ratios create the appearance of precision and objectivity. However, without measurement, this appearance is misleading. The numbers are not estimates of measurable attributes; they are constructed quantities derived from assumptions that are not themselves subject to validation. The comparison of interventions on this basis is therefore not a comparison of measures, but a comparison of constructed indices. The distinction is critical, yet it is not recognized within the knowledge base.

The fifth implication is institutional. The reproduction of this pattern across both Canadian and Australian knowledge bases indicates that it is not tied to specific organizational structures or policy environments. Academic centers, policy-linked institutes, and data-driven research units all exhibit the same profile. This suggests that the framework is transmitted through training, reinforced through publication, and aligned with policy requirements in a way that sustains its continuity. The system does not depend on coordination or agreement; it depends on reproduction. New entrants adopt the framework as given, and existing institutions reinforce it through established practices. Those progressing through the educational process are never introduced to the axioms of representational measurement let alone the importance of manifest as opposed to latent attributes. This is not deliberate; the concepts are entirely absent

The sixth implication is that incremental reform is unlikely to succeed. The problem is not that HTA methods are incomplete or require refinement. The problem is that the core constructs do not satisfy the conditions required for measurement. Improving model structure, increasing data quality, or refining parameter estimates does not address this issue. As long as the underlying constructs remain unchanged, the outputs will remain non-measurable. The framework cannot be corrected from within because the conditions required for correction, measurement and falsification, are not part of the system.

A seventh implication concerns responsibility. HTA informs decisions about the allocation of healthcare resources, the adoption of new technologies, and the prioritization of interventions. These decisions have direct consequences for patients and populations. When they are justified on the basis of numerical claims that do not meet the standards of measurement, there is a failure of analytical responsibility. This is not a matter of intent, but of structure. The system produces outputs that appear to support decisions, but those outputs are not grounded in measurable evidence. The obligation is therefore to admit to numerical storytelling.

The final implication is that a transition is required. The consistency of the results across jurisdictions indicates that the current framework is stable, but it is not scientifically valid. Moving forward requires a shift from numerical construction to representational measurement. Without this shift, HTA will continue to produce outputs that are internally coherent but externally indeterminate.

THERE IS AN EASY WAY OUT FOR THERAPY IMPACT ASSESSMENT

The question that now presents itself is disarmingly simple: is there an easy way out? After decades of analyses, thousands of publications, and the global institutionalization of cost-effectiveness frameworks across agencies and academic centers, the identification of a fundamental measurement failure appears, at first sight, to demand a radical and disruptive response. It is therefore natural to assume that the only path forward would require dismantling existing structures, to retrain entire cohorts of researchers, and abandoning established practices in favor of something unfamiliar and complex. That assumption is wrong. The difficulty does not lie in the absence of a solution. It lies in the reluctance to acknowledge that a solution already exists and is, in fact, straightforward.

The core of the issue is not methodological sophistication but measurement legitimacy. Once it is accepted that utilities, QALYs, and cost-effectiveness ratios do not meet the axioms of representational measurement, the continuation of their use as quantitative claims becomes untenable. But this does not leave a vacuum. It does not require the invention of a new science. The alternatives are already embedded within the structure of normal scientific practice. They have always been available, and they are conceptually simple.

There are, in essence, two pathways. The first is the use of manifest, directly observable attributes expressed as linear ratio measures. These include counts, durations, and other physical quantities: hospital days avoided, readmissions prevented, time to event, medication possession, clinic visits, and similar units. These are not abstractions. They are observable, reproducible, and possess the properties required for arithmetic operations. They have a true zero, support multiplication and division, and can be compared across contexts without ambiguity. In other words, they are already consistent with the axioms of measurement that govern all of normal science.

The second pathway addresses latent constructs; attributes that cannot be observed directly but are inferred from structured observations. Here, the solution is equally well established: Rasch measurement**,** yielding a logit ratio scale. This is not an exotic or speculative method. It is a formal measurement framework grounded in the same principles of invariance and dimensional consistency that underpin physical measurement. It provides a means to transform ordinal responses into a scale with interval and, when properly interpreted, ratio properties. The resulting logit scale allows for meaningful comparison, estimation of change, and evaluation of intervention impact. It does not require the summation of arbitrary scores or the imposition of preference weights. It requires adherence to a set of rules that ensure the existence of a measurable attribute.

These two approaches manifest ratio measures and Rasch logit ratio scales are not departures from established practice. They are aligned with it. They reflect the same commitment to measurement that defines physics, chemistry, and engineering. They do not require complex training beyond what is already expected of researchers operating in quantitative disciplines. They do not require the construction of elaborate simulation frameworks to compensate for the absence of data. They provide, instead, a direct and transparent basis for making claims that can be evaluated, replicated, and, crucially, falsified.

This is where the notion of an “easy way out” becomes meaningful. The transition does not require abandoning quantitative analysis. It requires abandoning the use of quantities that are not measures. It requires a shift in focus from constructed indices to measurable attributes. The tools are already available. The protocols can be specified. The data required for manifest measures are routinely collected. Instruments for latent constructs can be developed and validated within existing research programs supported by comprehensive software programs such as WINSTEPS that have been available for some 40 years.. The barrier is not technical. It is conceptual.

The real challenge, therefore, is not how to move forward, but how to acknowledge the need to do so. Admitting that a long-standing framework fails to meet the requirements of measurement is not a trivial step. It raises questions about prior analyses, published findings, and policy decisions that have relied on these constructs. It confronts institutions and individuals with the possibility that what has been treated as quantitative evidence is, in fact, numerical storytelling; representation without measurement. This is uncomfortable. It invites resistance.

Yet the alternative, continuing to use constructs that do not meet measurement standards is not neutral. It perpetuates a system in which claims cannot be empirically evaluated in the scientific sense. It maintains a framework that produces numbers without measurable attributes. Over time, this does not strengthen the field; it isolates it from the broader evolution of scientific knowledge.

But the cat is out of the measurement bag. The evidence is not only conclusive in demonstrating systematic measurement inversion; it is now being presented to a global HTA audience across academic centers, agencies, and policy domains. This is no longer an isolated critique or a methodological disagreement at the margins. It is a direct challenge to the foundations of the HTA knowledge base; the global memeplex of false measurement.

What makes the current moment different is that the issue can no longer be treated as implicit or peripheral. The interrogation of knowledge bases across multiple jurisdictions including academic centers in Canada, Australia, and now the United States demonstrates that the pattern is not contingent or context-specific. It is consistent, reproducible, and structural. The same probability–logit configurations emerge regardless of institutional setting, methodological emphasis, or policy alignment. This is not variation. It is invariance.

The question is therefore no longer whether measurement principles are occasionally overlooked or imperfectly applied. It is whether they are present at all in the constructs that define HTA practice. The answer, as these analyses demonstrate, is unequivocal: they are not. The core metrics utilities, QALYs, and cost-effectiveness ratios fail to meet the axioms of representational measurement. Yet they are treated as if they do.

This changes the terms of engagement. The issue is no longer one of refinement, adjustment, or improved modeling. It is one of recognition. The knowledge base that has been constructed over four decades rests on numerical claims that lack the properties required for measurement. That conclusion is now public, replicable, and increasingly difficult to ignore.

Given this, the path forward is both clear and manageable. It begins with a simple acknowledgment: that existing constructs such as utilities, QALYs, and cost-effectiveness ratios do not satisfy the axioms of representational measurement and therefore cannot support quantitative claims in the scientific sense. This acknowledgment does not require repudiating all prior work. It requires reinterpreting that work as part of a framework that lacks measurement, and therefore lacks the properties associated with measurable evidence.

The next step is equally straightforward: to adopt measurement-based claims. For manifest attributes, this means specifying outcomes in terms of observable, ratio-scale quantities. For latent constructs, it means developing instruments that satisfy Rasch requirements and reporting results on a logit scale. Each claim is then defined by a protocol: a target population, a measurable outcome, a timeframe, and a method for evaluation. These claims can be tested, replicated, and compared. They align with the principles of normal science.

There is no need for an extended period of theoretical debate or methodological refinement. The principles are established. The methods are available. The transition can begin with a single claim, a single study, a single decision framework. From there, it can expand.

Is there an easy way out? In practical terms, yes. The route is not obstructed by technical barriers or the absence of alternatives. It is obstructed by the reluctance to take the first step: to admit that the current framework does not meet the standards it claims to uphold. Once that step is taken, the rest follows naturally.

CONCLUSION WHY THE PATTERN PERSISTS: REPRODUCTION OF A CLOSED KNOWLEDGE BASE

The consistency of results across US research centers requires explanation. As in the Canadian and Australian reports, the interrogation reveals a stable pattern in which statements grounded in the axioms of measurement are weakly endorsed, while incompatible propositions are strongly reinforced. The key question is not whether this pattern exists, but why it persists across institutions that differ in size, location, and research focus. The answer does not lie in independent analytical judgment. It lies in the reproduction of a closed knowledge base.

The first element of this reproduction is entry through training. Researchers entering the HTA field do not begin with first principles of measurement. They are introduced to a set of accepted methods such as cost-effectiveness analysis, QALYs, and decision modeling, presented as established tools rather than as propositions requiring validation. Textbooks, graduate programs, and supervisory relationships transmit these methods as the foundation of the discipline. The question of whether the underlying constructs meet the requirements of measurement is not raised. As a result, new entrants adopt the framework as given. The conditions required for measurement are excluded at the point of entry.

The second element is methodological lock-in. Within HTA, certain constructs are not optional; they are required. The QALY, utility measurement, and the reference case define what counts as acceptable analysis. Research proposals, grant applications, and submissions to policy agencies are expected to conform to these templates. Deviation is not encouraged. This creates a closed methodological environment in which the same constructs are repeatedly applied. The interrogation results reflect this: strong endorsement of QALYs and cost-effectiveness claims is not the result of evaluation, but of compliance with established expectations.

The third element is reinforcement through publication. Academic journals and peer review processes play a central role in maintaining the framework. Studies that employ accepted HTA constructs are recognized as methodologically sound, while those that challenge these constructs face barriers to publication. The review process does not interrogate the measurement properties of utilities or QALYs; it assesses adherence to established methods. This creates a feedback loop. Researchers produce work that conforms to expectations, journals publish it, and the published literature reinforces the legitimacy of the framework. Over time, the framework becomes self-validating.

The fourth element is alignment with policy. In Australia, as in other jurisdictions, HTA research is closely linked to decision-making bodies such as the Pharmaceutical Benefits Advisory Committee. Policy requirements specify the use of cost-effectiveness analysis and QALYs as the basis for evaluation. Academic research aligns with these requirements to ensure relevance and impact. The result is a convergence between academic practice and policy expectations. The same constructs are used in both domains, reinforcing their status as the standard approach. The interrogation results show that this alignment does not extend to measurement principles. The constructs are accepted, but the conditions required to justify them are not examined.

A fifth element is the absence of falsification. In a scientific framework, claims are subject to empirical test. Measurement provides the basis for this process, allowing hypotheses to be confirmed or refuted. In HTA, the core constructs are not falsified. QALY-based claims and cost-effectiveness ratios are not tested against measurable attributes; they are outputs of models that incorporate assumptions about preferences, utilities, and disease progression. These assumptions cannot be subjected to measurement validation. As a result, the framework does not correct itself. It evolves through elaboration rather than through refutation.

These elements together define a system that reproduces itself. Training introduces the framework, methodological requirements enforce it, publication reinforces it, policy aligns with it, and the absence of falsification prevents its correction. The result is a knowledge base, a that is stable and internally coherent, but closed to the principles that would allow it to engage with measurement. The interrogation results are therefore not surprising. They are the expected outcome of a system in which the axioms of measurement are excluded from the outset.

It is important to emphasize that this reproduction does not require coordination or intent. Individual researchers operate hardwired within the framework as it is presented to them. The consistency of the results across US centers is not evidence of agreement in the sense of deliberate consensus. It is evidence that the same framework is being applied across domains. What appears as methodological alignment is, in fact, the replication of a shared doctrine.

The persistence of this pattern has implications beyond the US context. When the same results are observed across jurisdictions with different institutional arrangements, it suggests that the underlying framework is not local but global. The reproduction of measurement inversion is not confined to a particular country or set of institutions. It is a feature of HTA as a field.

The conclusion follows directly. The pattern persists because the framework that generates it is transmitted, reinforced, and protected from challenge. As long as the conditions required for measurement are not part of that framework, the pattern will continue. Change will not come from incremental adjustment within the existing system. It will require the introduction of measurement as a foundational requirement, and the abandonment of constructs that do not meet that requirement. Until then, the reproduction of measurement inversion will remain a defining feature of HTA research as exemplified by numerical storytelling.

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.

WEBSITE REFERENCES

UNITED STATES: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE SCHAEFFER CENTER FOR HEALTH POLICY AND ECONOMICS, UNIVERSITY OF SOUTHERN CALIFORNIA

UNITED STATES: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE CENTER FOR HEALTH DECISION SCIENCE, HARVARD UNIVERSITY

UNITED STATES: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE UNIVERSITY OF WASHINGTON CHOICE INSTITUTE

UNITED STATES: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE DUKE-MARGOLIS INSTITUTE FOR HEALTH POLICY, DUKE UNIVERSITY

UNITED STATES: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE LEONARD DAVIS INSTITUTE OF HEALTH ECONOMICS, UNIVERSITY OF PENNSYLVANIA

UNITED STATES: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE STANFORD UNIVERSITY HEALTH POLICY CENTER

UNITED STATES: THE ABSENCE OF REPRESENTATIONAL MEASUREMENT AND THE HEALTH ECONOMICS GROUP, BLOOMBERG SCHOOL OF PUBLIC HEALTH, JOHNS HOPKINS UNIVERSITY