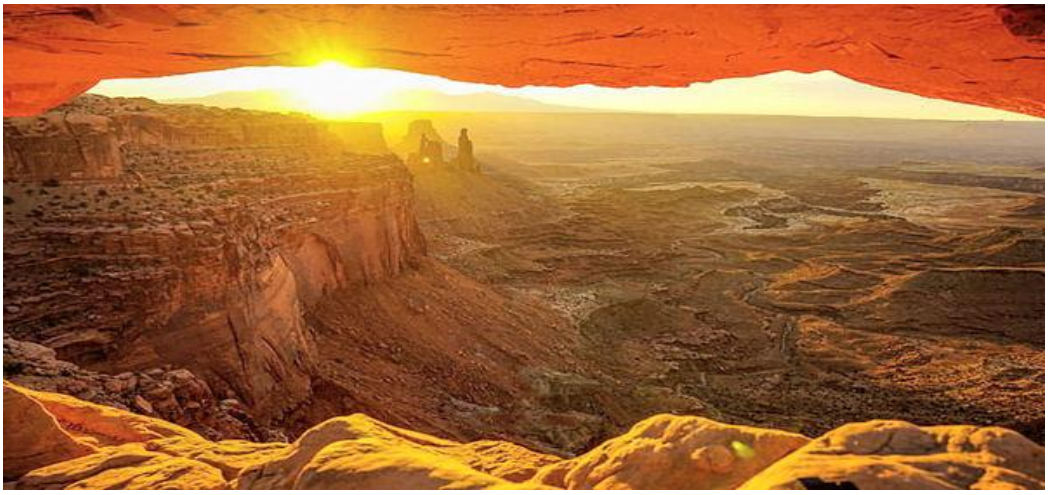
# MAIMON RESEARCH LLC

# ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



# REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

# UNITED KINGDOM: THE DEATH OF THE REFERENCE CASE IN HEALTH TECHNOLOGY ASSESSMENT

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

**LOGIT WORKING PAPER No 203 FEBRUARY 2026**

**www.maimonresearch.com**

**Tucson AZ**

## ABSTRACT

*This paper examines the reference case simulation model as the dominant analytical framework in health technology assessment (HTA) and argues that it has no legitimate scientific role. The central claim is not that reference case simulations are imperfect or overly simplified, but that they are structurally incapable of generating empirically evaluable, falsifiable claims about the real-world impact of competing therapies. Their persistence reflects administrative convenience and institutional consensus rather than epistemic validity.*

*The paper begins by tracing the origins of reference case simulations to an institutional need for comparability, standardization, and decisional closure under uncertainty. Simulation emerged as a procedural solution to heterogeneous evidence and long time horizons, not as a response to a measurement problem. This origin shaped a framework optimized for consistency and reproducibility rather than empirical confrontation. The requirements of scientific modelling are then set out as a fixed benchmark. Scientific models must generate testable implications, rely on quantities with admissible measurement properties, distinguish error from uncertainty, and expose claims to the risk of refutation. These requirements are non-negotiable conditions for cumulative knowledge.*

*Against this benchmark, the paper shows that reference case simulations fail decisively. Their outputs, most notably cost-per-QALY and cost-effectiveness ratios, are constructed from quantities that do not satisfy the axioms of representational measurement. Preference-based utilities lack unidimensionality and true zero properties, while cost aggregates are heterogeneous composites rather than measures of a single attribute. With the sole exception of time, the simulation model is assembled from non-measures, rendering its arithmetic inadmissible and its outputs non-empirical. The analysis further demonstrates why reference case simulations cannot generate falsifiable claims. Hypothetical cohorts, lifetime horizons, sensitivity analysis, and probabilistic modelling systematically insulate outputs from empirical exposure. Divergence from observed outcomes results in recalibration rather than refutation. Simulation thus replaces testing with procedural robustness.*

*A central contribution of the paper is the demonstration that reference case simulations function to protect multiattribute utilities. Utilities cannot survive direct empirical confrontation; simulation provides the environment in which they can be multiplied, aggregated, and projected without ever being observed. The survival of utilities and the survival of simulation are therefore inseparable.*

*Finally, the paper shows why repair is impossible. Methodological refinement, improved data, transparency, or expanded uncertainty analysis cannot correct failures that are architectural rather than technical. The reference case cannot be reformed into a scientific model without abandoning the features that define it.*

*The paper concludes that reference case simulation must be abandoned as a foundation for HTA. In its place, HTA must adopt a portfolio of single-attribute, empirically evaluable claims supported by lawful measurement: linear ratio measures for manifest outcomes and Rasch logit ratio*

*measures for latent traits. Only such a framework can restore falsification, learning, and scientific accountability to health technology assessment.*

*The Appendix situates the persistence of reference case simulation within a sociological framework. Drawing on critiques of the Strong Programme in the sociology of scientific knowledge, it explains how simulation-based HTA has shifted from evidence-based evaluation to consensus-driven legitimacy. The Appendix does not defend this position; it clarifies how the substitution of institutional agreement for empirical constraint helps explain the resilience of non-falsifiable modeling practices despite their failure to meet scientific standards.*

# INTRODUCTION

Reference case simulation models occupy a privileged and largely unquestioned position in health technology assessment (HTA). They are routinely presented as the integrative core of evidence-based decision making, capable of combining clinical data, quality-of-life estimates, costs, and assumptions into a coherent assessment of value over a lifetime horizon. Their authority rests not on direct empirical performance, but on methodological standardization, internal consistency, and adherence to prescribed "good practice." This paper argues that such authority is misplaced. Reference case simulations are not scientific models in the normal sense; they are epistemic closure devices that preclude falsification, insulate inadmissible quantities, and halt the evolution of objective knowledge.

The problem addressed here is not that reference case simulations are imperfect approximations of reality. All scientific models are approximations. The problem is that reference case simulations are structurally incapable of being wrong in any empirically meaningful sense. They do not generate testable predictions about observable therapy performance in health systems. Instead, they generate internally coherent narratives whose outputs change only when assumptions are altered. Disagreement is resolved procedurally, through sensitivity analysis, scenario selection, or consensus on parameter ranges, rather than empirically. In this respect, reference case simulations function as administrative instruments, not scientific hypotheses.

Historically, reference case models arose to solve an institutional problem: how to impose comparability and decisional finality across heterogeneous submissions. By mandating common assumptions, time horizons, discount rates, and outcome metrics, HTA agencies sought to standardize evaluation and avoid ad hoc decision making. This administrative motivation is understandable. What is not defensible is the subsequent reclassification of these simulations as sources of evidence. Standardization does not confer empirical meaning, and comparability does not substitute for testability.

The epistemic status of reference case simulations becomes especially problematic when viewed alongside the collapse of their primary inputs. As shown elsewhere, multiattribute utility instruments do not satisfy the axioms of representational measurement and cannot support lawful arithmetic. Reference case simulations do not correct this failure; they depend on it. Simulation allows non-measurement quantities to be multiplied, aggregated, and projected precisely because no empirical encounter is required. The model never meets the world. Outputs are shielded from refutation by design.

This insulation has profound consequences. In normal science, models are tools for learning because they expose claims to the risk of being wrong. Predictions fail, anomalies accumulate, and theories are revised or abandoned. Reference case simulations invert this logic. When outcomes diverge from expectation, the response is not refutation but recalibration. Error is absorbed into the model, not allowed to challenge it. Over time, this produces an expanding literature of numerical outputs with no epistemological standing beyond their compliance with methodological rules.

This paper therefore treats the reference case simulation as a failed scientific object. Its "death" is not rhetorical, but diagnostic. A framework that cannot generate falsifiable claims, cannot be empirically updated, and cannot distinguish error from uncertainty cannot serve as the quantitative foundation of HTA. What follows examines why reference case simulations arose, what scientific modeling requires, what these simulations actually do, why repair is impossible, and what must replace them if HTA is to recover its status as a field concerned with the real-world impact of competing therapies.

Central to this deconstruction are three elements that are central to measurement theory: Stevens' 1946 paper on scales of measurement with the key requirements of interval and ratio measures to support arithmetic [1]; (ii) the formalization by Krantz et al in 1971 of the axioms of representational measurement [2] and (iii) the Rasch model for transforming observations to interval and by transformation logit measures proposed in 1960 and formalized in representational measurement terms by Wright in 1977 [3] [4] .

## 1. WHY REFERENCE CASE SIMULATIONS WERE CREATED

Reference case simulation models were created to solve an institutional problem, not a scientific one. As health technology assessment expanded in scope during the late twentieth century, decision-making bodies faced increasing pressure to evaluate a growing volume of submissions across disparate disease areas, interventions, and evidentiary standards. The challenge was not how to test claims empirically in health systems, but how to impose order, comparability, and decisional closure on heterogeneous evidence within constrained timeframes. Reference case simulations emerged as a procedural solution to this administrative dilemma.

At their core, reference case models were designed to standardize evaluation. By specifying common assumptions, such as lifetime horizons, discount rates, outcome metrics, and perspectives, HTA agencies sought to ensure that submissions could be compared on a uniform basis. The "reference case" was intended to function as a methodological template rather than a hypothesis about the world. Its value lay in consistency, not truth. This distinction is crucial, because standardization is a bureaucratic virtue, whereas empirical validity is a scientific one.

Simulation was the natural vehicle for this standardization. Many of the outcomes of interest in HTA unfold over long time horizons and cannot be observed directly at the point of decision. Rather than restricting claims to what could be measured and tested, HTA adopted modeling as a way to extend short-term evidence into long-term projections. Simulation promised completeness where data were incomplete and continuity where observation was impossible. In doing so, it shifted the evidentiary burden away from empirical exposure and toward assumption management.

The rise of reference case simulations also reflected the institutional need for decisional finality. Health systems require decisions even when evidence is uncertain or incomplete. Reference case models provided a mechanism to close deliberation by producing a single, authoritative output, often a cost-effectiveness ratio, that could be compared against a threshold. This output did not need to be correct in an empirical sense; it needed to be defensible within an agreed methodological framework. Once that framework was accepted, disagreement could be managed procedurally rather than scientifically.

Importantly, the creation of reference case simulations coincided with the growing acceptance of composite outcome measures and preference-based utilities. Simulation allowed these quantities to be embedded within elaborate structures without requiring them to confront empirical reality. The model became the site where disparate elements, clinical data, utilities, costs, and assumptions, could coexist without ever being tested as a whole. In this sense, reference case simulations were not neutral analytical tools; they were enabling technologies for the use of quantities that could not otherwise sustain empirical scrutiny.

Early debates around modeling in HTA often acknowledged uncertainty, but they framed it as a technical challenge rather than an epistemic one. Uncertainty was to be explored through sensitivity analysis, probabilistic sampling, or scenario testing, not through empirical refutation. The model itself was never at risk. Instead, it was treated as a permanent scaffold to which new parameters could be attached. Over time, this approach became codified in guidelines and reinforced through publication norms, training programs, and professional incentives.

Crucially, reference case simulations were never intended to function as testable scientific models. They were intended to be acceptable, reproducible, and comparable. These are legitimate administrative goals, but they are not substitutes for empirical evaluation. By elevating procedural consistency over falsifiability, HTA institutionalized a form of modeling whose success was defined by adherence to rules rather than by performance in the world.

Understanding why reference case simulations were created clarifies why their failures are structural rather than accidental. They were built to manage uncertainty, not to resolve it; to close decisions, not to expose claims to risk. The sections that follow examine why this administrative solution cannot be reconciled with the requirements of scientific modeling and why the continued reliance on reference case simulations prevents HTA from functioning as a cumulative, empirically grounded discipline.

## 2. WHAT SCIENTIFIC MODELLING REQUIRES

Scientific modelling is not defined by complexity, computational sophistication, or methodological consensus. It is defined by its relationship to empirical reality. A scientific model is a conjectural representation of the world whose value lies in its capacity to be tested, to fail, and to be revised or abandoned in light of evidence. In other words, in Popper's term, to contribute to the evolution of objective knowledge [5] . Models do not acquire scientific standing by being plausible, elegant, or widely used; they acquire it by exposing their claims to the risk of being wrong.

At the most basic level, scientific models must generate testable implications. These implications need not be immediately observable, but they must be, in principle, confrontable with evidence. A model that cannot specify conditions under which it would be refuted is not a scientific hypothesis but a formal exercise. The central criterion is not uncertainty reduction but falsifiability. Uncertainty is ubiquitous in science; insulation from refutation is not.

Scientific modelling also requires a clear distinction between assumptions and claims. Assumptions are provisional starting points; claims are assertions about the world that stand or fall with evidence. In a scientific model, assumptions are exposed through predictions. If predictions fail, assumptions are revised or rejected. When assumptions are protected from empirical challenge, by redefining outcomes, extending horizons, or recalibrating parameters, the model ceases to function as a vehicle for learning.

A further requirement is empirical anchoring. Models must be anchored in quantities that possess empirical meaning. This does not imply that all model components must be directly observable, but it does require that quantities used in computation correspond to measurable attributes with known scale properties. Arithmetic performed on quantities that do not satisfy measurement axioms cannot yield meaningful results, regardless of how carefully the model is constructed. In scientific modelling, numerical manipulation is subordinate to measurement validity.

Time plays a critical role in scientific modelling. Models that project outcomes over time must specify how predictions will encounter observation. Short horizons, interim endpoints, and rolling evaluation are not methodological weaknesses; they are the mechanisms by which models remain vulnerable to refutation. A model that projects beyond any feasible observational window effectively removes itself from empirical accountability. Longevity of projection does not increase scientific value; it often diminishes it.

Scientific models also require error visibility. When predictions diverge from observed outcomes, that divergence must be interpretable as model failure rather than as an occasion for adjustment. In normal science, anomalies accumulate. They are not smoothed away through recalibration; they exert pressure on the model itself. The capacity to distinguish error from uncertainty is essential. Without that distinction, models cannot contribute to the evolution of objective knowledge.

Finally, scientific modelling is cumulative. Models are provisional structures that improve over time because they are constrained by evidence. Later models supersede earlier ones not because they are more comprehensive, but because they survive empirical challenge more successfully. Cumulative knowledge depends on selective retention: models that fail are discarded, and their failures inform successors. Where no model can fail decisively, no cumulative progress is possible.

These requirements are not aspirational ideals; they are the operational norms of normal science, recognized since the scientific revolution of the 17$^{th}$ century. They apply equally to theoretical physics and applied disciplines. A model that violates them may be useful for exploration, illustration, or decision support, but it cannot claim scientific authority. The purpose of modelling in science is not to produce definitive answers under uncertainty, but to structure conjectures in a way that allows the world to answer back.

The significance of these requirements becomes clear when contrasted with the structure of reference case simulations. The issue is not whether such simulations are carefully constructed, peer reviewed, or widely endorsed. The issue is whether they function as scientific models at all. The sections that follow examine that question directly by analyzing what reference case simulations actually do and why their design places them outside the tradition of empirical modelling.

## 3. THE REFERENCE CASE MODEL: WHAT IT DOES

Reference case simulation models do not function as scientific hypotheses about therapy performance in real health systems. They function as integrative accounting frameworks that convert heterogeneous inputs into standardized outputs under a prescribed set of assumptions. Their defining characteristic is not prediction, but closure. By design, the reference case model absorbs uncertainty, normalizes disagreement, and produces a single authoritative result that can support a decision.

The reference case begins by fixing a methodological frame: perspective, time horizon, discount rates, outcome measures, comparators, and costing conventions are specified in advance. These choices are not derived from empirical considerations but imposed for consistency. Once fixed, they define the boundaries of admissible analysis (e.g., a Markov framework). Claims that fall outside the frame are either reformulated to fit or excluded. The model therefore precedes the evidence rather than emerging from it.

Within this frame, the model assembles inputs from multiple sources: short-term clinical trials, observational studies, registry data, expert opinion, and, critically, preference-based utility scores. These inputs are not integrated through empirical testing but through structural assumption. Transition probabilities, state definitions, and parameter values are selected to ensure internal coherence, not empirical exposure. The resulting structure resembles a logical machine rather than an experimental one.

The primary output of the reference case is a counterfactual narrative. It describes what would happen to a hypothetical cohort over an extended, often lifetime horizon under alternative interventions. This narrative does not correspond to any observable population. Its purpose is not to be observed, but to be compared. The model constructs parallel worlds and evaluates them relative to each other. Because these worlds are not empirically accessible, their differences cannot be tested.

Uncertainty within the reference case is managed procedurally. Sensitivity analysis, probabilistic sampling, and scenario testing are used to demonstrate how results vary when assumptions are altered. These exercises are often presented as robustness checks, but they do not expose the model to refutation. They show that conclusions are stable within a range of assumptions, not that the assumptions themselves are wrong. The model remains intact regardless of outcome.

Crucially, reference case simulations lack failure conditions. There is no observation that, if realized, would invalidate the model as a model. Divergence between model outputs and real-world experience does not count as failure, because the model does not claim to predict the real

world. Instead, such divergence is interpreted as evidence of uncertainty, data limitations, or the need for recalibration. Error is reabsorbed into the framework rather than allowed to challenge it.

The reference case also plays a gatekeeping role. By specifying acceptable methods and outcomes, it disciplines submissions into a common format. This has practical advantages for agencies, but it also shifts evaluation from empirical performance to methodological compliance. A submission is judged not by whether its claims are borne out in practice, but by whether it conforms to the reference case. Scientific disagreement is replaced by procedural adjudication.

Perhaps most importantly, the reference case model transforms non-measurement into apparent measurement. Quantities that lack admissible scale properties such as preference-based utilities are rendered operational through simulation. The model permits multiplication, aggregation, and projection precisely because it never confronts empirical reality. Arithmetic proceeds unchallenged because there is no observational checkpoint at which its meaning could be questioned.

In this sense, the reference case model does exactly what it was designed to do. It produces standardized, comparable outputs that support decisional closure under uncertainty. What it does not do and cannot do is generate empirically testable claims about therapy impact. The model's success as an administrative tool is inseparable from its failure as a scientific one. Understanding this distinction is essential before asking whether such models can be repaired, reformed, or replaced.

## 4. SIMULATIONS CANNOT GENERATE FALSIFIABLE CLAIMS

Reference case simulations cannot generate falsifiable claims not only because of their structural insulation from observation, but because the quantities they produce are not empirically meaningful in the first place. The canonical output of such models such as cost per QALY or related cost-effectiveness ratios rests on arithmetic performed on quantities that lack the properties required for empirical testing. When neither the numerator nor the denominator constitutes a unidimensional ratio measure, the resulting ratio cannot be falsified because it does not represent a claim about the world. In measurement terms it is bankrupt.

Falsifiability requires that a quantitative claim correspond to an observable attribute whose magnitude could, in principle, be shown to differ from what is asserted. Reference case simulations fail this requirement at the most basic level. Their outputs are ratios of costs typically composite aggregates of heterogeneous resource inputs and QALYs, which are constructed from preference-weighted, multiattribute utility scores multiplied by time. Neither component satisfies the axioms of representational measurement. As a result, the ratio itself has no empirical referent.

The denominator is decisive. QALYs are not unidimensional ratio measures. The utility component lacks unidimensionality, lacks a true zero, permits negative values, and excludes lawful transformation from ordinal responses. Multiplying such quantities by time does not create a ratio measure; it compounds a category error. A denominator that does not represent an extensive attribute cannot anchor falsification. There is no observable quantity against which a QALY prediction could be compared, because QALYs do not exist as measurable entities in the world.

The numerator fares no better. Costs in reference case simulations are not measures of a single attribute but composites of diverse resource categories to include hospital days, procedures, medications, monitoring, and overhead, each with distinct dimensions. Aggregating these inputs into a monetary total does not produce a unidimensional ratio measure of resource use; it produces an accounting convenience. While monetary expenditures can be counted, they do not represent a homogeneous physical quantity whose ratios carry empirical meaning across contexts. As with QALYs, there is no single attribute whose magnitude could refute a modeled cost. It fails again on the simple requirement of ratio measurement of a single attribute.

Even if one were to assume, contrary to fact, that both costs and effects satisfied the requirements of ratio measurement, a cost-effectiveness ratio would still fail as a meaningful quantitative object because it lacks dimensional homogeneity. The numerator and denominator represent fundamentally different attributes with different dimensions. A ratio of dollars to health, time, or utility does not represent an extensive quantity whose variation corresponds to variation in a single empirical attribute. Such ratios do not admit interpretation as magnitudes in the world; they are rates constructed for comparison, not measures of anything possessed by patients or health systems. Dimensional homogeneity is a necessary condition for meaningful arithmetic, and its absence cannot be remedied by scaling, normalization, or modeling sophistication.

Because both numerator and denominator lack measurement validity, cost-effectiveness ratios are meaningless before simulation even begins. There is no observation that could show a modeled cost per QALY to be wrong, because there is no empirical counterpart to observe. Discrepancies between modeled outputs and real-world experience cannot count as refutation; they are necessarily reinterpreted as uncertainty, context dependence, or assumption sensitivity.

Simulation structure reinforces this insulation. Reference case models describe hypothetical cohorts evolving through artificial health states over extended horizons. These cohorts do not exist, and their trajectories cannot be observed. When real patients experience different outcomes, those outcomes do not contradict the model because the model does not claim to describe them. Instead, parameters are updated, utilities revised, or structures refined. Error is absorbed rather than exposed.

The decisive point is this: reference case simulations cannot generate falsifiable claims because they do not generate claims about measurable attributes. Cost per QALY is not an empirical proposition; it is a numerical artifact of inadmissible arithmetic performed within a closed modeling system. No increase in transparency, data richness, or computational sophistication can alter that fact.

Had the implications of Stevens' 1946 typology of measurement scales been understood or taken seriously, the proposal to construct a cost-per-QALY ratio as a basis for resource allocation would have been dismissed immediately. Stevens made explicit that arithmetic operations are conditional on scale type: ordinal scales do not support addition, interval scales do not support multiplication, and only ratio scales permit the formation of meaningful ratios. The QALY denominator fails these requirements at every level. It is neither unidimensional nor ratio-scaled, and it lacks a true zero. The numerator, composed of heterogeneous resource inputs aggregated monetarily, likewise fails to represent a single empirical attribute. A ratio formed from two non-ratio quantities cannot yield

a meaningful number. This is not a subtle methodological concern; it is elementary measurement law. That such ratios came to dominate HTA practice reflects not an unresolved scientific controversy but a foundational lapse in measurement literacy that was subsequently institutionalized rather than corrected.

The central implication of this analysis is difficult to avoid. If one had set out deliberately to design an evaluation framework for therapeutic claims that embodied the maximum number of violations of representational measurement, the result would be indistinguishable from simulated cost-outcomes claims as currently practiced in health technology assessment. Such claims combine non-unidimensional constructs, inadmissible arithmetic, absence of true zero properties, lack of dimensional homogeneity, and categorical insulation from empirical falsification within a single analytic object.

Simulated cost-effectiveness ratios do not merely contain measurement errors; they systematically accumulate them. Preference-based utilities substitute ordinal social judgments for measurement. Composite cost aggregates collapse heterogeneous resource categories into a single monetary figure without preserving a single empirical attribute. The ratio formed between these quantities lacks both measurement validity and dimensional coherence. These defects are then embedded within hypothetical cohorts and extended time horizons that preclude any direct confrontation with observed therapy performance.

What makes this framework especially revealing is that none of these failures is incidental. They are not implementation mistakes, data limitations, or methodological oversights. They are necessary design conditions for producing a single authoritative output in the absence of lawful measurement. The framework cannot be repaired without abandoning its defining purpose. Simulated cost-outcomes claims therefore do not fail despite their structure; they succeed because of it. Their authority rests precisely on the fact that they do not and cannot correspond to measurable attributes or testable claims about the world.

## 5. THE REFERENCE CASE PROTECTS UTILITIES

The reference case simulation does not merely incorporate multiattribute utilities as one input among many; it exists in large part to protect them. Preference-based utility scores cannot survive direct empirical confrontation because they do not represent measurable attributes possessed by patients. The reference case model supplies the epistemic environment in which such quantities can be used, multiplied, and projected without ever encountering the world. In this sense, simulation is not neutral infrastructure. It is a protective device.

Multiattribute utilities lack an empirical counterpart. No patient possesses a utility value in the way they possess survival time, symptom frequency, or functional capacity. Utilities are relational judgments derived from population preferences over hypothetical health states. They cannot be observed longitudinally, compared across individuals as quantities, or tested against clinical outcomes. If HTA were forced to ask whether a modeled utility gain actually occurred in practice, the question would be incoherent. The utility would fail immediately, not because it was poorly estimated, but because there is nothing to observe.

The reference case model removes the need for observation. Utilities are assigned to abstract health states rather than to patients. These states exist only within the model. Once embedded, utilities are multiplied by modeled time, discounted, and aggregated across hypothetical cohorts. At no point is a utility claim exposed to empirical risk. The model never requires a patient to experience 0.15 more utility units than another. The quantity functions entirely within the impossible arithmetic of the simulation.

With the sole exception of time, which alone possesses ratio properties, the reference case simulation is a smorgasbord of non-measures. It assembles preference scores, composite cost aggregates, artificial health states, and transition structures that do not satisfy the axioms of representational measurement and subjects them to arithmetic that would be impermissible outside the protective confines of the model. The model's coherence is therefore internal rather than empirical. Arithmetic proceeds not because the quantities warrant it, but because the simulation shields them from exposure.

Lifetime horizons are central to this protection. Short-term horizons would invite confrontation with observed outcomes, changes in symptoms, function, or resource use that could be measured and compared. Lifetime simulation ensures that the critical outcomes always lie beyond observation. By projecting far into the future, the model prevents utilities from being evaluated as claims about actual therapy response. The longer the horizon, the safer the utility.

Sensitivity analysis further insulates utilities from challenge. When utilities are questioned, the response is not to reject them as non-measures but to vary them within plausible ranges. This reframes a categorical measurement failure as parameter uncertainty. The utility itself is never at risk; only its assumed value is adjusted. Probabilistic sensitivity analysis intensifies this effect by embedding uncertainty into distributions, ensuring that any deviation from expectation can be absorbed as stochastic variation rather than treated as error.

The reference case also enforces the continued use of utilities by design. By mandating QALYs as the outcome of interest, cost-per-QALY ratios as the decision metric, and lifetime horizons as standard practice, HTA guidelines foreclose alternatives. Single-attribute outcomes, linear ratio measures of manifest events, and Rasch-based latent trait measures are treated as incomplete or supplementary. The framework does not merely tolerate utilities; it requires them. Escape routes are closed.

This protective role explains why multiattribute utilities and reference case simulations have persisted together despite repeated conceptual challenges. Utilities supply the numbers required for cost-effectiveness ratios; simulations supply the environment in which those numbers cannot be falsified. The relationship is symbiotic. Remove the simulation, and utilities are exposed as non-measures. Remove the utilities, and the simulation loses its central output.

Understanding this relationship is essential because it shows why reform efforts fail. Improving utility elicitation, refining valuation methods, or enhancing model transparency does not address the core problem. As long as utilities remain shielded from empirical exposure by simulation, they will continue to function as authoritative quantities despite lacking measurement validity.

The reference case therefore does more than organize evidence; it stabilizes a system of pseudo-quantification. By protecting utilities from the risk of being wrong, it converts preference judgments into permanent parameters and transforms non-measurement into apparent evidence. Any attempt to restore HTA as a science concerned with real-world therapy impact must dismantle this protective structure. Without that step, utilities and the simulations that depend on them will persist not because they measure anything, but because they cannot be tested.

## 6. WHY REPAIR IS IMPOSSIBLE

The failures of the reference case simulation are not technical defects that can be corrected. They are structural consequences of the model's purpose and design. Repair is impossible because the reference case was never intended to function as a scientific model capable of empirical refutation. It was designed to deliver decisional closure under uncertainty, and every feature that enables that function simultaneously disables scientific accountability and any claim to meeting the axioms of representational measurement.

One proposed route to repair is methodological refinement: better data inputs, richer evidence synthesis, improved transition structures, or more realistic assumptions. These refinements leave the core problem untouched. A simulation built from non-measures does not become empirically meaningful by being populated with more precise estimates. Precision applied to quantities that lack admissible scale properties only produces more precise error. No refinement can transform preference-based utilities or composite cost aggregates into unidimensional ratio measures, and without such measures, arithmetic remains inadmissible.

Another proposed repair is expanded uncertainty analysis. Probabilistic sensitivity analysis, value-of-information analysis, and Bayesian updating are often presented as ways to strengthen inference. In fact, they weaken it. These techniques embed uncertainty into the model so thoroughly that empirical divergence can no longer count as failure. When every outcome is framed as a realization of uncertainty, no outcome can refute the model. Bayesian updating does not rescue falsifiability; it replaces error with posterior belief. A framework that can absorb any observation without rejection cannot learn.

Transparency is also offered as a remedy. Making assumptions explicit, documenting model structure, and publishing code may improve reproducibility, but reproducibility is not validity. A transparent non-falsifiable model remains non-falsifiable. Clarity about assumptions does not expose them to empirical risk; it merely legitimizes them procedurally. Scientific models are not validated by disclosure but by confrontation with evidence.

Shortening time horizons is sometimes suggested as a compromise. Yet partial repair is not repair. As long as the model's central outputs remain impossible cost-per-QALY ratios built from non-measures, the horizon length is irrelevant. Even over short periods, QALYs cannot be observed, and composite costs cannot be interpreted as measures of a single attribute. Shortening the projection does not create an empirical referent where none exists.

The most revealing evidence that repair is impossible lies in the categorical exclusion of alternatives. A genuinely repairable framework would be open to replacement of invalid

components. The reference case is not. Rasch-based latent trait measures, single-attribute ratio outcomes, and protocol-driven empirical evaluation are consistently marginalized as "incomplete" or "non-comprehensive." This is not accidental. Admitting such measures would dismantle the reference case's integrative function. Repair would require abandonment of the very features that define the model.

The failures are therefore overdetermined. The reference case violates measurement axioms, performs inadmissible arithmetic, precludes falsification, substitutes calibration for refutation, and enforces its own continuation through guidelines and publication norms. Each failure independently disqualifies it as a scientific model. Together, they make reform conceptually incoherent. There is no version of the reference case that can be both what it is and what science requires.

This conclusion is uncomfortable because it implies that decades of methodological effort were misdirected. Yet the alternative, to continue refining a framework that cannot generate testable claims, is worse. Science progresses by abandoning structures that cannot fail. The reference case simulation cannot fail, because it does not make claims about measurable attributes in the world. Its outputs can be replaced, updated, or superseded, but never refuted.

Repair is therefore not an option. The choice is between dismissing the reference case and reconstructing HTA around empirically evaluable, single-attribute claims, or continuing a system of numerical storytelling insulated from evidence. Only the former allows HTA to function as a discipline concerned with the real-world impact of competing therapies.

## CONCLUSION

This paper has shown that the reference case simulation is incompatible with the requirements of scientific inquiry in health technology assessment. Its failure is not a matter of imperfect implementation, data limitations, or methodological immaturity. It is a structural failure rooted in the model's purpose and design. Reference case simulations were created to impose comparability and decisional closure, not to generate empirically testable claims about therapy performance. That administrative function cannot be reconciled with the demands of normal science.

At the center of this failure is arithmetic performed on non-measures. With the sole exception of time, the reference case model is constructed from quantities that do not satisfy the axioms of representational measurement. Preference-based utilities lack unidimensionality and true zero properties; composite cost aggregates do not represent a single empirical attribute; modeled health states are classificatory devices rather than measurable quantities. Ratios formed from such components, most notably cost per QALY, cannot be falsified because they do not correspond to anything observable in the world.

Simulation does not correct this defect; it conceals it. By embedding non-measures within abstract structures, projecting them over lifetime horizons, and treating assumptions as parameters rather than claims, the reference case insulates its outputs from empirical risk. Sensitivity analysis substitutes for falsification, calibration replaces refutation, and disagreement is resolved

procedurally rather than scientifically. The result is not cumulative knowledge but numerical storytelling with the same cast of characters.

An unsettling implication of this analysis is the collective silence that has surrounded these failures. There has been no sustained, concerted effort within HTA to confront the manifest deficiencies of the reference case framework, despite their visibility once elementary principles of representational measurement are applied. Thousands of HTA practitioners, methodologists, reviewers, and educators have remained effectively mute in the face of violations that would be disqualifying in any other quantitative discipline. It is one thing to overlook an isolated error of measurement or to debate a contested assumption. It is quite another to accommodate, indeed to institutionalize, an analytical framework that constitutes an archetype of measurement failure, combining non-unidimensional constructs, inadmissible arithmetic, lack of dimensional homogeneity, and insulation from falsification in a single object. This accommodation cannot be explained as inadvertence or technical disagreement. It reflects a deeper normalization of epistemic compromise, in which methodological convenience and procedural consensus have displaced the obligation to ensure that quantitative claims actually measure something in the world.

This persistence is best explained by institutional inertia rather than epistemic legitimacy. Once codified in guidelines and reinforced through publication norms, the reference case became self-protecting. Alternatives that would expose claims to empirical testing, single-attribute outcomes, ratio measures of manifest events, and Rasch-based latent trait measures, are systematically marginalized because they threaten the integrative fiction on which the reference case depends. In this environment, methodological conformity is rewarded, while challenges grounded in measurement theory are treated as peripheral or impractical. The result is a framework that cannot fail decisively and therefore cannot be corrected.

The implications are stark. If health technology assessment is to remain a discipline concerned with the real-world impact of competing therapies, it must abandon the reference case simulation as a foundation for value claims. In its place must be a portfolio of single-attribute, empirically evaluable claims supported by lawful measurement: linear ratio measures for manifest outcomes and Rasch logit ratio measures for latent traits, evaluated prospectively and revised in light of evidence.

If this reconstruction does not occur, HTA will remain, as it has for 40 years, a monument to a memeplex of irrelevant pseudo-measurement. It remains a closed system of procedural compliance and model-based authority, disconnected from patient experience and clinical reality. Only by dismantling the reference case and restoring falsification as an operational principle can HTA claim scientific accountability, meet its duty of care to patients and clinicians, and enter the tradition of normal science in which claims advance precisely because they are exposed to the risk of being wrong.

# APPENDIX

## FROM EVIDENCE TO CONSENSUS: THE COLLAPSE OF SIMULATION AND HEALTH TECHNOLOGY ASSESSMENT AS A SOCIOLOGICAL BELIEF SYSTEM

The persistence of reference case simulation in health technology assessment cannot be explained by its empirical success, because it does not generate empirically testable claims. Nor can it be explained by its measurement validity, because its core quantities violate the axioms of representational measurement. The only remaining explanation is sociological. Reference case simulation persists because it has become a socially stabilized belief system; one that substitutes consensus for evidence and procedural agreement for empirical warrant.

This transformation marks a fundamental shift in the epistemic character of HTA. In normal science, evidence precedes agreement. Claims are advanced, tested, refuted, or provisionally retained based on their performance against observation. Consensus, when it emerges, is an outcome of empirical success, not a prerequisite for legitimacy. In contrast, the reference case framework reverses this order. Agreement on methods, assumptions, and outputs is treated as sufficient for credibility, while empirical exposure is deferred indefinitely or rendered irrelevant.

Simulation plays a central role in this reversal. By relocating claims from the world to the model, reference case simulations remove the need for direct confrontation with evidence. Outputs are no longer propositions about observable therapy performance but artifacts of an agreed analytical process. Their authority derives not from correspondence with reality, but from adherence to methodological conventions codified in guidelines, journals, and professional norms. Once this shift occurs, debate is no longer about whether a claim is true, but about whether it conforms.

This is precisely the condition described, though not endorsed, by sociological accounts of scientific knowledge that emphasize consensus, institutional power, and boundary maintenance, notably under the strong program [6]. David Wootton's critique of the strong program in the sociology of scientific knowledge is particularly instructive for understanding how reference case simulation has come to dominate HTA despite its failure to meet elementary scientific standards [7]. Wootton argues that the strong program does not merely offer a sociological explanation of how scientific beliefs arise; it advances a relativist epistemology in which the very notion of discovery is displaced. In this view, evidence is never uncovered as an independent constraint imposed by the world. Instead, it is constructed within specific social communities, shaped by their interests, norms, and power structures. What counts as "evidence" is therefore contingent, not objective.

In Wootton's reading, success within the strong program framework is judged not by correspondence with reality, but by a claim's ability to mobilize assent within a community. A theory succeeds because it persuades, not because it is true. Scientific knowledge becomes indistinguishable from other forms of social belief: its authority rests on rhetoric, institutional endorsement, and professional consensus rather than on empirical vulnerability. Disagreement is resolved through negotiation and stabilization, not through refutation. Crucially, there is no privileged role for reality to play as an arbiter. The HA memeplex does not recognize reality.

This position marks a sharp break with the tradition of science as an enterprise aimed at discovering how the world works. In Wootton's critique, the strong program replaces that tradition with an account of science as a cultural activity whose outputs are validated internally. Evidence does not confront belief; belief defines evidence. The very idea that a claim could be wrong independently of social recognition is rejected as naïve realism.

Applied to HTA, this relativist stance provides a disturbing but illuminating lens. Reference case simulation fits the strong program's conception of knowledge production almost perfectly. Model outputs are accepted not because they correspond to observable therapy effects, but because they conform to an agreed methodological template. Their authority derives from consensus, codified in guidelines, reinforced by journals, and maintained by professional networks. Challenges grounded in measurement theory or falsifiability fail not because they are incorrect, but because they relieve standards the system no longer acknowledges.

From this perspective, the persistence of cost-per-QALY modeling is not a puzzle but an inevitability. If evidence is socially constructed, then the absence of unidimensional measures, true zeroes, or empirical refutation is irrelevant. What matters is whether the modeling framework continues to coordinate decision making and command assent. The reference case succeeds because it stabilizes belief, not because it reveals anything about therapy impact.

Wootton's critique exposes the cost of this position. Once science is reduced to rhetoric, persuasion, and authority, it loses its distinguishing feature: the capacity to be corrected by reality. Claims can no longer fail decisively; they can only lose favor. Progress becomes indistinguishable from fashion. In such a framework, the accumulation of publications signals not the growth of knowledge, but the expansion of a belief system to a progression of reference model claims that may extend for decades.

The relevance to HTA is stark. Defending reference case simulation on sociological grounds, consensus, acceptability, transparency, is not a fallback justification; it is an admission that HTA has abandoned science as a way of coming to grips with reality. Wootton's deconstruction makes clear that this move does not rescue the enterprise. It reclassifies it. HTA becomes an object of sociological study rather than a source of objective knowledge about the effects of competing therapies.

In this sense, the strong program can explain the endurance of simulation-based HTA, but only by conceding its epistemic collapse. Once evidence is understood as constructed and success as persuasion, the very idea of falsifiable, measurement-grounded claims disappears. What remains is authority without truth, a condition that may sustain institutions and religions, but cannot sustain science.

However, what may be descriptively accurate as sociology becomes fatal when adopted implicitly as epistemology. When consensus replaces evidence as the criterion of validity, measurement axioms lose their force. Violations of unidimensionality, scale type, and falsifiability no longer disqualify claims; they are absorbed as features of accepted practice. Simulation outputs cannot be wrong in any decisive sense, because their legitimacy does not depend on being right. They persist as long as the community continues to recognize them.

This sociological stabilization explains the otherwise puzzling resilience of cost-per-QALY modeling in the face of repeated conceptual critiques. Challenges grounded in measurement theory, arithmetic admissibility, or empirical testability fail to gain traction because they appeal to standards the system no longer recognizes. The reference case does not respond to refutation; it responds to dissent by procedural accommodation. Sensitivity analysis replaces testing. Recalibration replaces rejection. Learning is redefined as refinement within a closed framework.

The result is a mature belief system with the outward appearance of science but without its internal discipline. Numbers circulate, models proliferate, and publications accumulate, yet none of these activities expose claims to the risk of failure. The system becomes self-referential. Its success is measured by replication of method, not by correspondence with reality. Evidence is no longer something that constrains belief; belief determines what counts as evidence.

Understanding HTA in these terms is not an act of rhetorical excess. It is the logical consequence of a framework that has abandoned falsification while retaining the language of quantification. Once reference case simulation is defended on the grounds of consensus, transparency, or acceptability rather than empirical performance, HTA has already crossed the boundary from science to sociology.

The choice is no longer between alternative modeling approaches, but between two fundamentally different conceptions of HTA: one as a scientific enterprise grounded in measurement and falsification, the other as a memeplex a socially stabilized system of numerical belief. In practical terms, the HTA memeplex now behaves as a closed cognitive framework that automatically excludes measurement-valid alternatives. Claims grounded in linear ratio or Rasch logit ratio scales are not debated or rejected; they simply fail to register as legitimate within the evaluative system. The situation recalls earlier moments in the history of science in which empirically grounded claims were not refuted but remained literally unrecognized; an epistemic blindness rather than a reasoned disagreement.

## ACKNOWLEDGEMENT

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

## REFERENCES

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

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

[3] Rasch G, Probabilistic Models for some Intelligence and Attainment Tests. Chicago: University of Chicago Press, 1980 [An edited version of the original 1960 publication]

[4] Wright B. Solving measurement problems with the Rasch Model. *J Educational Measurement.* 1977;14(2):97-116

[5] Popper K. *Objective Knowledge: An Evolutionary Approach*. Rev. ed. Oxford: Clarendon Press, 1979

[6] Bloor D. *Knowledge and Social Imagery*. 2nd ed. Chicago: University of Chicago Press, 1991

[7] Wootton D. The Invention of Science: A New History of the Scientific Revolution. New York: Harper Collins, 2015