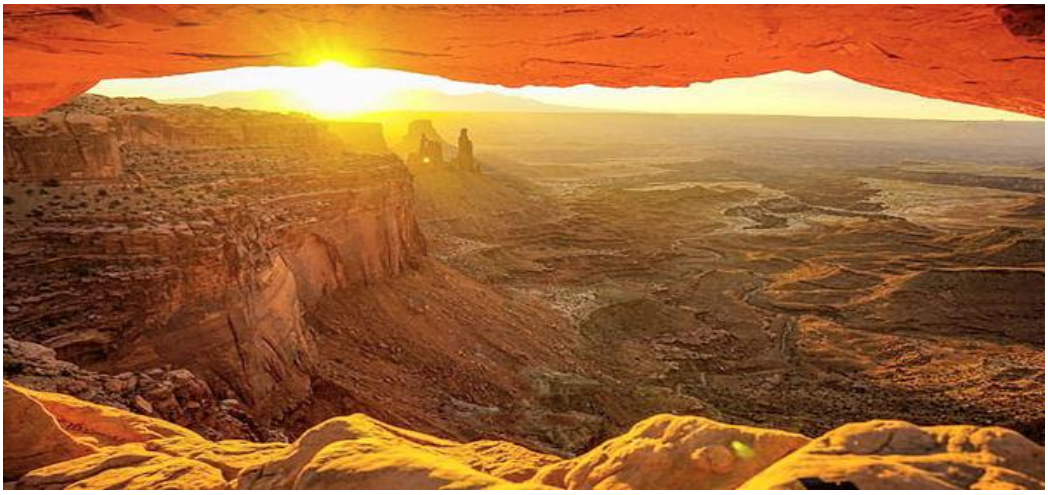
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: NONSENSE ON STILTS - NICE AND THE ABANDONMENT OF REPRESENTATIONAL MEASUREMENT

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

LOGIT WORKING PAPER No 27 JANUARY 2026

www.maimonresearch.com

Tucson AZ

NONSENSE ON STILTS: NICE AND THE ABANDONMENT OF SCIENTIFIC MEASUREMENT

HOW ADMINISTRATIVE DIKTAT WASTED FORTY YEARS

Jeremy Bentham famously dismissed empty abstraction as *"nonsense upon stilts"* claims elevated by authority rather than grounded in demonstrable meaning. That warning sits at the heart of the modern scientific enterprise. From the founding of the Royal Society in 1660, the guiding principle *nullius in verba,* take nobody's word for it, marked a decisive break from scholasticism, authority, and metaphysical assertion. Knowledge would no longer be validated by institutional decree or theoretical elegance, but by demonstrable structure, empirical discipline, and testable claims. Two centuries later, Stevens' formalization of measurement theory made that principle operational: numbers do not confer meaning by their presence alone; only when empirical attributes satisfy definable axioms may arithmetic be applied. Measurement precedes calculation. Without that ordering, numerical claims are not provisional science; they are category errors.

The NICE reference case overturned this entire lineage. It institutionalized arithmetic without measurement, authority without falsification, and consensus without empirical grounding. Where Newton insisted *hypotheses non fingo,* I feign no hypotheses, NICE requires precisely the opposite: hypothetical lifetime models populated by constructed utilities, aggregated preference scores, and threshold comparisons untethered from measurable quantities. These outputs are not tested against reality; they are negotiated within administrative frameworks designed to achieve closure rather than truth. In doing so, NICE replaced the scientific method with managed plausibility. It did not merely misunderstand measurement; it rendered measurement irrelevant. What emerged was not evaluation in the tradition of normal science, but a bureaucratic scholasticism, numerical in appearance, metaphysical in substance, whose authority derives not from empirical refutation, but from institutional repetition. Bentham would have recognized it immediately. Newton would have rejected it outright. And the Royal Society's founders would have seen in it the very form of knowledge they created modern science to escape.

EXECUTIVE SUMMARY

This Logit Working Paper examines the emergence, institutionalization, and global diffusion of health technology assessment (HTA) as a system of numerical storytelling rather than scientific evaluation. It argues that the contemporary HTA framework, centered on reference-case modeling, utility-based outcomes, and cost-effectiveness thresholds, represents a fundamental departure from the standards that govern scientific inquiry.

Section I situates HTA within the broader context of the Scientific Revolution, emphasizing that modern science advanced by rejecting scholastic authority in favor of empirically testable claims. Measurement, falsification, and replication were not optional refinements but the structural foundations that enabled objective knowledge to evolve. Quantification without demonstrable measurement status was explicitly rejected as pre-scientific reasoning.

Section II traces the evolution of objective knowledge through the principles of representational measurement theory. It shows that arithmetic operations are meaningful only when numbers represent attributes with defined structural properties. From these axioms follow two and only two admissible forms of measurement: linear ratio scales for manifest attributes and Rasch logit ratio scales for latent traits. These principles establish why measurement must precede arithmetic and why subjective outcomes cannot be quantified through scoring conventions alone.

Section III examines the emergence of NICE as a modern form of scholasticism. Rather than subjecting therapy claims to falsifiable testing, NICE institutionalized reference-case modeling as the primary evaluative framework. This shifted assessment away from empirical confrontation toward internally coherent simulation, thereby replacing provisional scientific claims with administratively convenient closure.

Section IV demonstrates how this framework has blocked the future of HTA. By eliminating falsification, preventing cumulative learning, and normalizing arithmetic without measurement, the NICE model created a global memeplex that reproduces method without discovery. The result is a discipline rich in procedure yet incapable of scientific progress.

The paper concludes that meaningful reform requires abandoning reference-case decision models and re-establishing measurement as the gatekeeping condition for evidence. Without this transition, HTA cannot evolve as a scientific enterprise.

SECTION I: THE SCIENTIFIC REVOLUTION: MEASUREMENT, FALSIFICATION, AND THE END OF AUTHORITY

The scientific revolution did not begin with equations, instruments, or data. It began with a rejection. What was abandoned was not ignorance, but authority as a source of truth. For centuries, knowledge had been organized around commentary, interpretation, and appeal to recognized masters. Claims were judged by coherence with accepted doctrine rather than by exposure to empirical failure. The revolution that emerged in Europe over the sixteenth and seventeenth centuries overturned this epistemic order. It replaced authority with measurement, plausibility with falsification, and narrative explanation with empirical risk [1].

The founding insight was deceptively simple: numbers have no meaning unless they represent something real. Mathematical elegance, internal consistency, and logical plausibility are insufficient to establish truth. A claim about the world must be structured so that it can fail. Without the possibility of failure, there is no discovery, only belief. This was the defining break from scholasticism. Science was not born from better arguments, but from rules that determine when arguments cease to matter.

This transformation was institutionalized most clearly in the ethos of the Royal Society, captured in its enduring motto *nullius in verba* "take nobody's word for it." The phrase did not mean that expertise was irrelevant. It meant that no authority, however prestigious, could substitute for demonstrable evidence. Claims were to be accepted provisionally, tested publicly, and discarded when contradicted. Knowledge would advance not through consensus, but through structured disagreement with reality itself.

Measurement was central to this transformation. It was not an optional technical refinement; it was the mechanism that made falsification possible. Science demands measurement. Only when an attribute could be expressed as a quantity with stable properties could competing claims be meaningfully compared. Measurement provided the invariant units that allowed experiments to be replicated across time, place, and investigators. Without measurement, replication degenerates into repetition; different studies using different scales, producing numbers that cannot be aligned, tested, or accumulated [2].

The early natural sciences understood this with remarkable clarity. Time, distance, mass, and later energy were not merely recorded; they were defined in ways that preserved constant relations. The success of physics did not arise from mathematical sophistication alone. It arose because physical attributes possessed structures that permitted ratio measurement: equal units and meaningful zero points. Arithmetic followed measurement, not the reverse. Calculations were permitted only because the underlying quantities had already earned the right to be calculated.

This ordering, measurement first, arithmetic second, is not a convention. It is a logical necessity [3]. Arithmetic operations are meaningful only when the scale on which numbers reside permits them. Addition requires equal units. Multiplication requires a true zero. Without these properties, numerical manipulation produces symbols that look quantitative but are not measures. This distinction, formalized in representational measurement theory in the twentieth century, was already implicitly understood centuries earlier. The pioneers of science did not perform arithmetic

on categories, rankings, or preferences. They knew that numbers without structure were numerology.

The scientific revolution therefore imposed discipline on mathematics itself. Mathematics was not rejected; it was constrained. It could describe reality only when anchored to empirical structure. Detached from that structure, it reverted to speculation. The triumph of science was not that it became mathematical, but that it learned when mathematics was allowed to speak.

This insight underlies the evolution of objective knowledge. Knowledge grows when claims are framed so that they can be corrected. A hypothesis is proposed, tested, refined, or discarded. Progress emerges through successive approximations toward truth, not through definitive closure. Crucially, this process depends on quantities that remain stable under comparison. Without invariant measures, disagreement cannot be resolved empirically. Competing claims become incommensurable narratives rather than rival explanations of the same phenomenon.

The logic of falsification makes this unavoidable [4]. A claim can be falsified only if its outcome can be measured in the same units under comparable conditions. If each study uses a different scale, or if the scale lacks defined properties, failure cannot be distinguished from reinterpretation. Error becomes ambiguous. Science stalls not because data are lacking, but because meaning dissolves.

This is why measurement is inseparable from accountability. When quantities are measured, claims can be held to account. When they are merely scored, indexed, or modeled, accountability evaporates. Numbers remain, but their connection to reality becomes negotiable. The discipline that makes science uncomfortable, its willingness to be wrong, depends entirely on the discipline of measurement.

These principles apply with equal force beyond the physical sciences. When inquiry moves into the human domain, the challenge intensifies rather than diminishes. Attributes such as functioning, symptom burden, need fulfillment, or patient experience are not directly observable. They are latent. Yet the requirement for measurement does not disappear. If anything, it becomes more stringent. Latent attributes cannot be assumed into existence by assigning numbers to responses. They must be constructed through formal measurement models capable of producing invariant units.

Here again, the logic of science admits no shortcuts. Subjective observations do not become quantitative because they are collected systematically, analyzed statistically, or averaged across samples. Ordinal data remain ordinal regardless of how many decimal places are attached to them. Statistical association cannot create measurement where structure is absent. Correlation does not manufacture quantity.

The development of Rasch measurement in the mid-twentieth century represents the extension of the scientific measurement project into the human sciences [5]. Rasch models do not score responses; they test whether observed data conform to the requirements of measurement. They impose unidimensionality, examine item invariance, and produce a logit ratio scale that expresses the possession of a latent attribute in constant units. Where the data fail these requirements,

measurement is rejected. This is not a limitation of Rasch; it is its virtue. Measurement is earned, not presumed.

The significance of this framework cannot be overstated. It provides the only scientifically defensible route by which subjective observations can be transformed into quantities suitable for arithmetic. Without it, latent attributes remain descriptive categories. With it, they become measurable constructs capable of supporting falsifiable claims. The logic mirrors that of the scientific revolution itself: impose constraints first, permit calculation only afterward. The Rasch framework for latent traits is entirely consistent with the axioms of representational measurement [6].

What unites all of this is a single ordering principle: measurement precedes arithmetic. This is not philosophy in the abstract. It is the rule that determines whether numbers participate in science or merely decorate argument. When this ordering is respected, knowledge can evolve. When it is inverted, numbers proliferate while understanding stagnates. The tragedy that follows when this principle is abandoned is not immediately obvious. Systems can appear sophisticated. Models can be complex. Outputs can be precise. Yet precision without measurement is illusion. It produces certainty without truth and closure without discovery. The very features that make such systems attractive to administrators, definitive results, apparent comparability, clean thresholds, are the features that insulate them from correction.

The scientific revolution taught a harder lesson. There is no final closure. Claims remain provisional. Evidence accumulates slowly. Error is not eliminated; it is managed through exposure to refutation. Any framework that promises certainty without measurement is not advancing science; it is retreating from it. This is the standard against which all contemporary evaluative frameworks must be judged. Not by their computational elegance, not by their institutional acceptance, and not by their policy convenience, but by their fidelity to the principles that made science possible in the first place. When measurement is bypassed, arithmetic becomes authority. And when arithmetic becomes authority, the scientific revolution is undone—not by ignorance, but by forgetting why its rules existed at all.

SECTION II: THE EVOLUTION OF OBJECTIVE KNOWLEDGE: WHY SCIENCE ADVANCES ONLY THROUGH MEASURABLE ERROR

Objective knowledge does not grow by accumulation. It grows by correction. This distinction, often misunderstood, lies at the heart of modern science. Knowledge advances not because more facts are collected, but because existing claims are exposed to structured risk of failure and are revised or abandoned when they fail. The central mechanism is not confirmation but falsification. A theory survives not because it is proven true, but because it has not yet been shown false under conditions where falsity would be unmistakable. This is impossible in the HTA memeplex.

For this process to function, claims must be framed in a form that permits refutation. They must make contact with the world through measurable consequences. If outcomes cannot be measured in invariant units, then disagreement cannot be resolved empirically. Competing interpretations persist indefinitely, not because reality is ambiguous, but because the instruments of adjudication are absent. This is why the evolution of objective knowledge is inseparable from measurement.

Without measurement, there is no stable reference point against which claims can be tested. Evidence becomes interpretive rather than decisive. Error becomes rhetorical rather than empirical. Scientific debate turns inward, circulating within a closed system of assumptions instead of confronting the world.

Karl Popper's insight was not that theories must be falsifiable in principle, but that they must be falsifiable in practice. A claim that cannot be shown wrong by conceivable observation is not scientific, regardless of how plausible or widely accepted it may be. But falsification is impossible without measurement. Observation alone is insufficient. What matters is not that something is seen, but that it can be compared quantitatively with what was predicted.

This requirement immediately excludes vast classes of numerical outputs from scientific status. Numbers derived from ordinal categories, composite indices, preference rankings, or weighted scores cannot support falsification because they lack fixed meaning. When a result differs from expectation, one cannot know whether the theory failed, the scale shifted, or the interpretation changed. The error signal is blurred. The system cannot learn. The HTA memeplex can never learn.

Scientific progress therefore depends on the existence of invariant quantities. These are not quantities that remain constant, but quantities whose units remain stable across contexts. When a unit changes meaning across populations, instruments, or time, comparison collapses. A difference of "five units" no longer signifies the same thing. Without invariant units, replication becomes symbolic. Studies can be repeated endlessly without converging on truth.

This is the fatal weakness of knowledge systems built on scores rather than measures. Scores are contingent. They depend on item selection, weighting schemes, response formats, and scoring rules. Change any of these and the number changes meaning. Statistical sophistication cannot rescue this instability. Regression, significance testing, and sensitivity analysis all assume that the underlying variable is quantitative. They do not create the requirement,

Objective knowledge evolves only when claims are anchored to quantities that remain meaningful under repeated testing. This is why the physical sciences progressed so rapidly once measurement systems stabilized. Once time, distance, and mass were measured consistently, competing theories could be tested decisively. Errors accumulated visibly. Incorrect theories were discarded. Progress followed not from agreement, but from elimination.

The human sciences face a more difficult problem, but not a different one. Attributes such as health status, functioning, or quality of life are not directly observable. They cannot be weighed or timed. Yet the logic of objective knowledge does not relax in their presence. If claims about these attributes are to evolve, they must be framed in a way that allows empirical contradiction.

Here the distinction between latent constructs and latent measures becomes decisive. A latent construct is a theoretical idea. A latent measure is a quantified representation that satisfies measurement axioms. Confusing the two is catastrophic. Merely naming a construct does not make it measurable. Attaching numbers to responses does not transform those numbers into quantities.

Rasch measurement represents the necessary bridge between theory and empirical test for latent attributes. It does not assume that a latent variable exists because an instrument has been designed. It tests whether responses conform to the requirements of a single underlying dimension with invariant item functioning. Only when those conditions are met does a measure emerge. When they are not met, the claim collapses.

This discipline is what enables objective knowledge to evolve in domains involving subjective data. Without it, claims cannot be falsified. One can always reinterpret the scale, revise the scoring algorithm, or redefine the construct. Error never lands cleanly. Knowledge stagnates under a veneer of activity.

The contrast with systems that reject this discipline is stark. When falsification is replaced by plausibility, knowledge becomes administrative. Claims are accepted because they appear reasonable, align with precedent, or satisfy institutional needs. Disagreement is managed through negotiation rather than resolution. Competing models coexist indefinitely, not because they are equally valid, but because there is no mechanism to eliminate them.

This is precisely how belief systems persist. When error cannot be demonstrated, beliefs become immune to correction; hence the HTA memeplex. They are refined, extended, and elaborated, but never overturned. This is not scientific evolution; it is doctrinal stability. Thomas Kuhn described this condition as normal science operating within a paradigm [7]. But Kuhn also recognized that paradigms collapse when anomalies accumulate that cannot be absorbed. Crucially, anomalies can accumulate only when measurements reveal inconsistency. Without measurement, anomalies dissolve into interpretation. Paradigms persist not because they explain the world, but because nothing can prove them wrong.

This explains why certain evaluative frameworks can dominate for decades without producing genuine knowledge growth. They generate outputs, reports, and recommendations, yet no cumulative learning occurs. The same debates recur. The same assumptions are recycled. The same disputes remain unresolved. What evolves is technique, not understanding.

When a system such as HTA prioritizes closure over correction, it abandons the evolution of objective knowkedge, Decisions may still be made, but they are insulated from learning. Once a claim is accepted, it is rarely revisited. The system moves forward not by testing its assumptions, but by embedding them deeper. Objective knowledge, by contrast, is ruthless. It discards failure. It refuses closure. It treats every claim as provisional. It requires that claims be expressed in forms that allow the world to push back. This is uncomfortable, slow, and politically inconvenient. But it is the only mechanism that separates science from administration.

This is where the connection to HTA becomes unavoidable. Any framework that aspires to evaluate therapy impact must be capable of learning from experience. It must be able to say, after implementation, whether its prior claims were wrong. If it cannot do so, it is not assessing technology; it is classifying it. A system built on non-measures cannot support this learning. Outcomes cannot be compared across time. Claims cannot be reproduced in invariant units. Deviations can always be attributed to context, heterogeneity, or model uncertainty. Nothing ever fails decisively. The system becomes epistemically closed. This is not a failure of intent. It is a

failure of structure. Without measurement, falsification cannot occur. Without falsification, objective knowledge cannot evolve. Without evolution, the system becomes static, regardless of how dynamic it appears.

The lesson of the scientific revolution is therefore not historical ornament. It is a warning. Any framework that reintroduces authority, consensus, or plausibility as substitutes for measurable error is not modern science in a new form. It is pre-scientific reasoning with modern mathematics; the HTA memeplex. The tragedy is not that such systems exist. The tragedy is that they can flourish while presenting themselves as scientific. When numbers are mistaken for measures, arithmetic for evidence, and models for knowledge, the appearance of rigor masks the absence of learning. The evolution of objective knowledge demands something far more austere. It demands humility before measurement. It demands acceptance of provisional claims. And above all, it demands the willingness to be wrong in ways that cannot be negotiated away.

This is the standard against which all evaluative institutions must ultimately be judged. Not by how persuasive their outputs appear, nor by how widely they are adopted, but by whether their claims can fail in the world. Where failure is impossible, knowledge cannot grow. And where knowledge cannot grow, science has quietly ended; even if its language remains everywhere in use.

SECTION III: NICE AND THE RETURN OF SCHOLASTICISM: NUMERICAL AUTHORITY WITHOUT MEASUREMENT

By the late twentieth century, the intellectual conditions for scientific evaluation were well established. The axioms of measurement were known. The distinction between ordinal, interval, and ratio scales had been formalized. The role of falsification in the evolution of objective knowledge had been articulated and absorbed across the sciences. Against this backdrop, the emergence of the National Institute for Health and Care Excellence (NICE) did not represent an advance in evaluative method. It represented a retreat.

NICE did not merely fail to apply the lessons of the scientific revolution; it inverted them: where science demands that measurement precede arithmetic, NICE institutionalized arithmetic without measurement; where science requires that claims be provisional and exposed to empirical refutation, NICE substituted model closure for falsification; where scientific knowledge advances through error correction, NICE constructed a framework designed to prevent error from ever being observed. This is why the correct analogy for NICE is not failed science, but scholasticism.

Medieval scholasticism did not reject reason. It elevated it; but within limits. Logic flourished, but only inside a closed doctrinal system. Debate was permitted so long as it did not challenge foundational premises. Authority determined admissibility, not empirical confrontation with reality. Arguments were judged internally coherent, not externally testable. NICE replicates this structure with modern symbols. The reference case functions as doctrine. It defines what counts as acceptable evidence before any empirical question is asked. Utilities must be used. QALYs must be constructed. Lifetime models must be applied. Thresholds must be referenced. These requirements are not conclusions drawn from measurement; they are axioms imposed in advance of it. Once accepted, all further reasoning occurs within this closed universe.

The NICE reference case does not ask whether utilities are measures. It assumes they are. It does not ask whether QALYs are dimensionally coherent. It presumes they are. It does not test whether multiplication of time by preference weights is meaningful. It mandates it. The result is an evaluative system that begins not with observation, but with metaphysics disguised as methodology. This is scholastic reasoning in quantitative dress.

The hallmark of scholasticism is not error but immunity to error. A system is scholastic when its core claims cannot be falsified because the rules of evaluation guarantee their survival. NICE achieved precisely this outcome by relocating evaluation from the world to the model. In the reference case, claims are never tested against outcomes. They are tested against assumptions. Sensitivity analysis becomes the highest form of scrutiny, yet sensitivity analysis only explores how conclusions change when beliefs change. It does not confront the model with reality. It cannot reveal error in the strong sense because there is no empirical benchmark to contradict. The world does not speak back.

Instead, NICE produces what appears to be rigorous analysis: structured models, probabilistic distributions, confidence intervals, scenario analyses. Yet none of these operations address the foundational question of whether the quantities being manipulated are measures at all. Arithmetic proceeds as ritual rather than inference. The reference case thus becomes a self-sealing system. If results are unfavorable, assumptions are adjusted. If uncertainty is high, further modeling is recommended. If disagreement persists, deliberation replaces testing. At no point is a claim exposed to refutation by observed outcomes measured in invariant units. This is not an accident. It is the purpose.

NICE emerged in an environment where policymakers required closure. Decisions had to be made at launch, not revisited indefinitely. Budgets demanded predictability. Manufacturers demanded access. Health systems demanded administrative certainty. A framework grounded in falsification could not deliver this. Claims subject to ongoing empirical challenge would never settle. The reference case solved this problem elegantly. It offered apparent scientific rigor while eliminating the possibility of empirical embarrassment. Once a model had been constructed and reviewed, the decision was complete. Future outcomes could always be explained away as contextual variation rather than failure of the original claim. Closure replaced learning. Administrative closure and convenience replaced any commitment to the evolution of objective knowledge for therapy impacts

This is the defining scholastic feature. Scholastic systems are optimized not for truth discovery, but for decision finality. They prioritize coherence, stability, and authority over correction. They generate elaborate reasoning structures that appear intellectually impressive while remaining detached from empirical consequence. NICE's fixation on the QALY embodies this logic perfectly.

The QALY is not a measure. It violates unidimensionality, lacks a true zero, permits negative values while claiming ratio properties, and combines fundamentally different attributes through multiplication. These failures are not subtle. They are elementary. Yet the QALY became the centerpiece of NICE not because it satisfied measurement axioms, but because it provided a single number capable of supporting administrative decisions. The elegance of one number outweighed

the incoherence of its construction. Once installed, the QALY acquired authority not through validation, but through repetition. It appeared in guidelines, models, submissions, and academic papers. Its ubiquity became its justification. This is another hallmark of scholasticism: authority emerges from tradition rather than empirical warrant.

Over time, the distinction between measurement and convention disappeared. Utilities were treated as cardinal by decree. Thresholds acquired moral significance. Debates focused on parameter values rather than on whether the structure itself was admissible. Methodological energy was spent refining assumptions instead of questioning foundations. NICE thus created an evaluative culture in which asking whether something is measurable became improper. This is why the NICE framework proved so transmissible internationally. It did not demand measurement expertise. It did not require engagement with representational measurement theory. It did not require understanding of falsification or invariance. All that was required was compliance with a recipe. This made it extraordinarily attractive to jurisdictions seeking legitimacy without epistemic burden.

Countries adopting the NICE model could claim scientific sophistication simply by reproducing its forms. They could generate ICERs, apply thresholds, and conduct probabilistic sensitivity analyses without confronting whether any of these operations were meaningful. The appearance of science traveled faster than science itself ever could. The NICE memeplex did not spread because it was correct. It spread because it was administratively efficient.

Universities amplified this process. Teaching reference-case modeling was easier than teaching measurement theory. Students could learn to operate software without grappling with axioms. Journals could publish results without adjudicating scale validity. Reviewers could evaluate structure without understanding measurement. The scholastic order reproduced itself. In this environment, the Royal Society's motto, *nullius in verba*, take nobody's word for it, was quietly inverted. The reference case asks precisely the opposite: take the model's word for it. Accept its assumptions. Trust its structure. Believe its outputs. The world need not be consulted.

This is the most profound indictment of NICE's intellectual legacy. It did not merely introduce a flawed metric. It normalized a way of reasoning that makes empirical refutation unnecessary. It transformed evaluation from a scientific activity into a ritualized performance of quantification. Once that transformation occurred, the evolution of objective knowledge ceased. There could be no accumulation of error signals. No convergence toward better understanding of therapy impact. No learning across time. Each new technology would be evaluated afresh within the same closed framework, generating a new set of numbers immune to contradiction by experience.

Forty years later, the consequences are visible. The same debates recur endlessly. Models become more complex, but conclusions remain untestable. Disagreements persist because nothing can be settled empirically; angels and pins.. The system appears busy yet intellectually static. This is not science stalled; it is science displaced. NICE did not fail because it made a mistake. It failed because it replaced the logic of scientific inquiry with the logic of scholastic closure. It institutionalized numerical storytelling as governance. And in doing so, it provided the blueprint for a global HTA memeplex that speaks the language of evidence while insulating itself from the

discipline that evidence requires. If any evidence is required, we need look no further than the CHEERS 2022 guidance for imaginary reference case clams [8] [9] .

SECTION IV: NICE HAS BLOCKED THE FUTURE: HOW NUMERICAL STORYTELLING REPLACED SCIENTIFIC PROGRESS

The most damaging consequence of the NICE framework is not that it introduced an invalid metric, nor even that it institutionalized arithmetic without measurement. Its true legacy is more profound: NICE has blocked the future of health technology assessment by foreclosing the possibility of scientific progress itself. Science advances through error correction. Claims are made provisionally, exposed to empirical testing, and revised or abandoned when they fail. Measurement provides the medium through which this process occurs. Without invariant quantities, failure cannot be observed, and without observable failure, objective knowledge cannot evolve. The NICE reference case was constructed precisely to prevent this process.

By relocating evaluation from empirical outcomes to simulated projections, NICE severed the link between claims and reality. Therapy impact was no longer something to be observed and tested, but something to be inferred from a model. Once inference replaced confrontation, falsification disappeared. A claim could no longer be wrong in the strong scientific sense; it could only be sensitive, uncertain, or scenario-dependent. This is not an incidental flaw. It is the defining feature of the system.

In a reference-case world, a therapy is never shown to fail. If real-world outcomes diverge from projections, the divergence is attributed to context, implementation, adherence, population differences, or parameter uncertainty. The original claim remains untouched because it was never framed as falsifiable to begin with. It existed only inside a conditional model. The result is epistemic paralysis.

No HTA body operating under the NICE paradigm can learn systematically from experience. There is no mechanism by which initial claims are confirmed or refuted. There is no cumulative archive of measured outcomes. There is no evolving understanding of therapy impact grounded in invariant units. Instead, there is endless re-modeling. Each new submission restarts the ritual. Each new technology generates another simulation. Each decision closes the file without opening the future. This is why the system has remained unchanged for four decades. Not because it has succeeded, but because it cannot fail.

Contrast this with what a measurement-based HTA system would require. Claims would be single-attribute, unidimensional, and explicitly stated. Manifest claims would be expressed on linear ratio scales. Latent trait claims would be measured on Rasch logit ratio scales. Each claim would be accompanied by a protocol specifying population, timeframe, comparator, and outcome. Results would be observable. Claims could be reproduced or refuted. Such a system would be uncomfortable. It would never be closed. Decisions would be provisional. Evidence would accumulate. Errors would be visible. Knowledge would evolve. NICE was designed to avoid precisely this discomfort.

In administrative terms, this avoidance was understandable. Policymakers sought closure at launch. Budgets demanded certainty. Manufacturers wanted predictable access. A framework based on falsification would have required continuous reassessment and the admission that earlier decisions might be wrong. That burden was deemed unacceptable. So a different path was chosen. Instead of measurement before arithmetic, arithmetic was permitted first. Instead of falsifiable claims, plausibility sufficed. Instead of empirical testing, internal coherence was enough. Instead of scientific humility, confidence intervals were offered.

The cost of this choice was the future. Once the reference case became entrenched, any alternative framework grounded in measurement appeared alien. Single-claim evaluation looked simplistic compared to complex models. Rasch measurement seemed obscure compared to utility algorithms. Falsification looked naïve in a world accustomed to probabilistic storytelling. The NICE memeplex protected itself. Journals aligned their expectations accordingly. Training programs taught modeling rather than measurement. Agencies harmonized around NICE-style frameworks. Consultants optimized software tools for simulation. Over time, entire professional identities formed around practices that could not survive contact with representational measurement theory. At that point, reform ceased to be merely technical. It became existential.

To accept that measurement must precede arithmetic would be to admit that decades of published evaluations cannot support the claims made for them. To accept that QALYs are not measures would be to undermine thresholds, benchmarks, and comparative rankings. To accept Rasch as the sole route to latent trait measurement would invalidate vast libraries of patient-reported outcome research. The system therefore did not merely resist change; it rendered change unthinkable. This is how scientific revolutions are prevented; not by argument, but by institutional design. NICE did not need to suppress critics overtly. It simply created a framework in which critics had no place to stand. Measurement theory was not rejected; it was made irrelevant. Falsification was not denied; it was redefined. Evidence was not questioned; it was simulated. The language of science remained, but its logic was gone.

The global transmission of this framework now appears less mysterious. NICE offered something irresistible: a way to appear scientific without the burden of science. Countries could claim rigor while avoiding empirical risk. Institutions could produce numbers without confronting whether those numbers measured anything. In this sense, NICE was not merely influential. It was catalytic. What spread was not a method, but a permission structure; the permission to do arithmetic without measurement, to make claims without falsification, to substitute coherence for truth. That permission reshaped HTA worldwide.

The tragedy is that none of this was necessary. The principles of measurement were available. Stevens' scale typology had been published half a century earlier. Representational measurement theory was mature. Rasch measurement had already demonstrated how latent attributes could be quantified legitimately. Popper had articulated falsification as the engine of knowledge growth.

NICE ignored all of it. Not because the knowledge was unavailable, but because it was inconvenient. And once ignored at the founding moment, it became impossible to recover later. Foundational choices harden quickly. Frameworks become curricula. Curricula become

professions. Professions become gatekeepers. By the time the incoherence was recognized, the system was too large to question.

This is why the present moment matters. The accumulated evidence, now visible through systematic Artificial Intelligence large language model (LLM) interrogation makes clear how the error ridden basis of the NICE reference model. LLM logit diagnostics across agencies, journals, academic centers, and health systems, shows the same pattern everywhere: near-floor endorsement of measurement axioms; near-ceiling endorsement of arithmetic fictions. The profile is not local. It is structural. That structure traces back, historically and conceptually to NICE. This does not make NICE uniquely culpable. But it does make NICE uniquely responsible.

The institution that claimed authority to define value for money in healthcare also defined what would count as knowledge. In doing so, it redirected an entire field away from the scientific path and toward a numerically ornate but epistemically barren alternative. The consequence is forty years of effort without accumulation. Models have grown more complex. Data have grown larger. Software has grown more powerful. Yet no stable quantitative understanding of therapy impact has emerged. Debates recur unchanged. Disagreements persist unresolved. Evidence remains perpetually provisional, not because science demands humility, but because the system cannot test itself. This is the hallmark of blocked science.

CONCLUSION

The NICE reference case did not fail because it was imperfectly implemented, insufficiently refined, or applied with inadequate data. It failed because it was deliberately constructed on foundations that precluded scientific evaluation from the outset. Its central error was not technical but epistemic: the deliberate abandonment of representational measurement as the gatekeeping condition for quantitative claims.

By the time NICE formalized its framework in the late 1990s, the principles governing measurement were neither obscure nor contested. Stevens' typology of scales, published in 1946, had long established that arithmetic operations are permissible only when numbers possess the structural properties required to represent empirical attributes. This was not a philosophical preference but a logical constraint. Multiplication, aggregation, and ratio comparison cannot be justified on ordinal or composite scales, regardless of institutional endorsement or analytical convenience.

Yet the NICE reference case proceeded as if this knowledge did not exist; a lack which characterized most of the NOCE audience. Health state descriptions were converted into preference scores, ordinal responses were treated as if they possessed equal intervals, and products of time and utility were declared ratio measures by convention rather than demonstration. In doing so, NICE institutionalized arithmetic without measurement, replacing scientific inquiry with administratively satisfying calculation.

The consequence was decisive. A framework built on non-measures cannot generate falsifiable claims. It cannot support replication in the strong sense. It cannot accumulate objective knowledge over time. What it can produce are stable narratives, internally coherent, methodologically

elaborate, and immune to empirical refutation. The appearance of rigor masks the absence of measurement. This is why the reference case cannot be repaired. No refinement of modeling technique, no expansion of sensitivity analysis, and no recalibration of thresholds can correct a framework that never established the legitimacy of its numbers. The failure is structural, not incremental. The NICE reference case did not represent a flawed science awaiting correction. It represented a departure from science itself. Until measurement precedes arithmetic, HTA cannot claim to evaluate therapy impact. It can only tell numerical stories.

To unblock the future requires more than technical reform. It requires abandonment. The reference case must be relinquished as a decision variable. Composite constructs must be disallowed as claims. Measurement must be reinstated as a gatekeeping condition. Claims must be single, explicit, and evaluable. Latent attributes must be measured, not scored. Arithmetic must follow measurement or not occur at all. This is not a return to simplicity. It is a return to science. Until that transition occurs, health technology assessment will remain what NICE made it: a discipline fluent in numbers, rich in procedure, and incapable of learning. The greatest cost of numerical storytelling is not mispricing or access delay. It is the loss of discovery. When a system cannot tell when it is wrong, it cannot ever become right. NICE did not merely shape HTA. It defined its limits. The task now is to move beyond them.

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

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

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

[4] Popper, Karl R. Objective Knowledge: An Evolutionary Approach. Revised edition. Oxford: Clarendon Press, 1979.

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

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

[7] Kuhn, T. The Structure of Scientific Revolutions. 2nd Ed. Chicago: University of Chicago Press, 1970.

[8] Husereau, D, Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) 2022 Statement: Updated Reporting Guidance for Health Economic Evaluations. *Value in Health*. 2022; 25 (1): 3-9.

[9] Husereau, D, Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) 2022 Explanation and Elaboration: A Report of the ISPOR CHEERS Good Practices Task Force. *Value in Health.* 2022; 25 (1): 10-31.