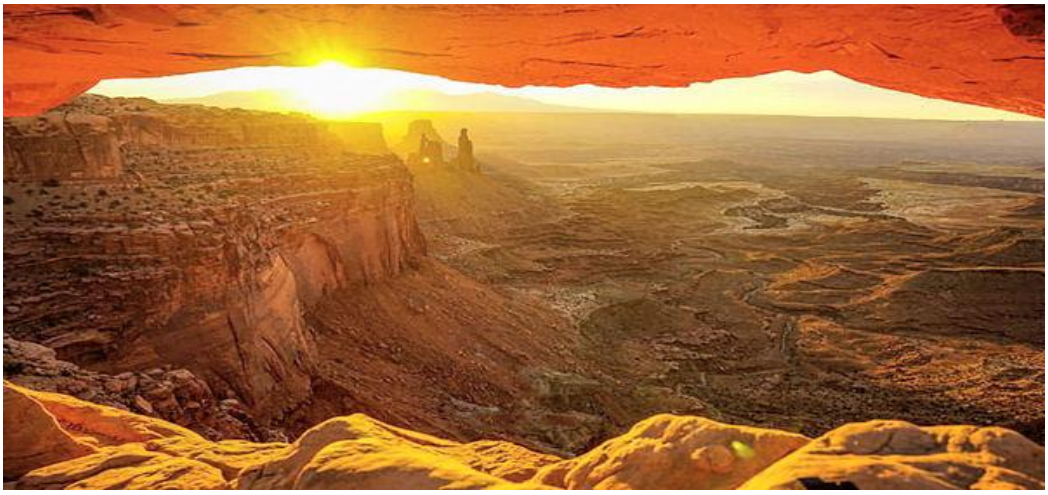
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: THE GLOBAL ODYSSEY OF THE NICE REFERENCE CASE

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

LOGIT WORKING PAPER No 28 JANUARY 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

The global dominance of the NICE reference case in health technology assessment (HTA) represents one of the most rapid and complete episodes of methodological convergence in the history of applied policy science. Within less than two decades, a nationally administrative pricing framework developed in England was adopted across Europe, North America, Asia-Pacific, and beyond. This diffusion occurred despite the framework's reliance on constructs, utilities, QALYs, preference-weighted health states, and long-horizon simulation models, that fail the axioms of representational measurement. The central question addressed in this paper is therefore not whether the reference case is scientifically defensible, a matter already resolved in the negative, but how a mathematically incoherent framework achieved such extraordinary global acceptance with almost no sustained resistance.

The analysis argues that the reference case did not spread as a scientific theory, but as an institutional technology. Its success lay not in empirical validation, falsification, or replication, but in its ability to deliver administrative closure under conditions of limited data, political pressure, and budget constraint. By providing standardized procedures, thresholds, and model outputs, the reference case enabled pricing and access decisions to appear evidence-based without requiring measurable claims. Adoption therefore occurred through imitation rather than scientific contestation.

The paper traces the mechanisms of transmission through UK academic centers, professional training pipelines, and international societies, most notably ISPOR, which normalized the framework as "best practice" while never requiring justification of its measurement foundations. The result was the formation of a closed epistemic loop in which teaching, journals, guidelines, and databases mutually reinforced a belief system whose core assumptions were never examined.

Critically, the reference case encountered little resistance because the audiences receiving it lacked the conceptual tools necessary to interrogate it. Training in representational measurement theory, scale-type admissibility, and the distinction between ordering and measuring had largely disappeared from economics, outcomes research, and HTA curricula. In this epistemic vacuum, numerical outputs were treated as quantitative by default, and falsification was quietly replaced by sensitivity analysis.

Using insights derived from large language model (LLM) diagnostics applied across national HTA corpora, the paper demonstrates that global convergence reflects not consensus but shared absence: the systematic exclusion of measurement as a gatekeeping requirement. The findings reframe the NICE reference case not as a flawed scientific paradigm, but as a durable administrative memeplex, one that prioritized procedural legitimacy over epistemic legitimacy. The paper concludes that restoring scientific credibility to HTA requires abandoning the reference case architecture and re-establishing measurement as a non-negotiable precondition for arithmetic, comparison, and policy inference.

THE GLOBAL TRANSMISSION OF THE REFERENCE CASE

The remarkable feature of the NICE reference case is not that it was constructed without measurement foundations. That failure has already been established. What demands explanation is something far more puzzling: how a framework built on mathematically indefensible claims achieved global acceptance with extraordinary speed and almost no resistance. Within less than two decades, a nationally administrative pricing tool developed in England became the dominant evaluative template across Europe, North America, Asia-Pacific, and beyond. Countries with entirely different health systems, legal traditions, and institutional cultures adopted the same core architecture: preference-weighted health states, QALYs, reference-case models, and threshold-based decision rules. This occurred not through coercion, treaty, or formal mandate, but through imitation. The reference case became not merely a method, but a norm.

This rapid diffusion cannot be explained by scientific merit. Scientific ideas spread slowly, through replication, contestation, and refinement. They face resistance. They generate counter-programs. They fracture into competing schools. None of this occurred with the reference case. Instead, the model propagated almost frictionlessly. Its adoption resembled administrative convergence, not scientific evolution. The key to understanding this lies in recognizing that the reference case was never transmitted as a scientific theory. It was transmitted as an institutional solution. Policy-makers across health systems faced a common problem: how to justify pricing and access decisions under conditions of limited data, political pressure, and budget constraint. They required a framework that could generate apparent objectivity without requiring empirical closure. The reference case provided exactly that. It did not ask whether claims were true. It asked whether assumptions were reasonable. It did not demand falsification. It delivered decisions.

Once NICE established the appearance of procedural legitimacy, a standardized model, a defined threshold, a reproducible template, the framework became exportable. It could be adopted without deep epistemic understanding because epistemic understanding was unnecessary to its operation. What mattered was not whether the numbers measured anything, but whether the process appeared rigorous. This explains the extraordinary role of UK academic centers in the transmission process. York, Sheffield, Oxford, and affiliated institutions became global training hubs. Students arrived from health ministries and universities worldwide, learned the NICE framework as "best practice," and returned home carrying not a theory but a template. What was exported was not the evolution of objective knowledge in the Popperian sense, but procedural competence for the reference case as an unchanging analytical framework [1].

The reference case thus spread through educational throughput rather than scientific validation. Its authority derived from institutional prestige, not empirical survival. The UK's long-standing reputation in economics, public administration, and applied policy analysis provided the credibility bridge. If this is how the UK does it, it must be right.

Critically, this transmission bypassed the normal mechanisms of scientific challenge. Journals did not demand measurement justification because journals had already internalized the same framework. Reviewers were trained within it. Editors had built careers on it. A closed epistemic loop formed, in which conformity replaced scrutiny.

This is how a memeplex operates. A memeplex does not survive because it is true; it survives because it is mutually reinforcing. Teaching reproduces methods. Journals publish compliant work. Agencies cite published work. Databases archive it. Guidelines reference the archive. Each component validates the others, and none can easily step outside without threatening its own legitimacy. The reference case was ideally suited for such a system. It was computationally sophisticated, rhetorically scientific, and epistemically lightweight. It required no measurement theory, no falsification protocols, no longitudinal replication. It replaced the uncertainty of science with the comfort of scenario analysis. It offered closure where science offers provisional acceptance.

Most importantly, it solved an administrative problem. Health systems do not want perpetual uncertainty. They want decisions. A framework that delivers a number, any number, that can justify a price decision is institutionally attractive. The reference case provided a defensible fiction: decisions could be framed as evidence-based even when the evidence was model-generated. This also explains why resistance never coalesced. Those best positioned to object to include health economists, outcomes researchers, HTA professionals were the very people whose professional identity depended on the framework's survival. To challenge the measurement basis of the QALY was not to propose an improvement; it was to question the legitimacy of an entire career structure. As a result, silence replaced critique. Stevens' scale typology was ignored [2]. Rasch measurement was excluded with silence on the role of the axioms of representational measurement [3,4,5]. Falsification was redefined as sensitivity analysis. The Royal Society's motto, *nullius in verba*, was inverted. Rather than trust no one's word, the reference case asked decision-makers to trust the model and not the established falsification standards of science.

Over time, the framework became naturalized. New entrants encountered it not as a contested idea but as settled doctrine. Textbooks taught it. Software encoded it. Agencies mandated it. What began as a national administrative expedient hardened into global orthodoxy. The speed of this transmission is therefore not mysterious once the nature of what was transmitted is understood. This was not science traveling. It was governance technology. It was not a theory competing for truth, but a mechanism competing for adoption. And mechanisms that promise closure, comparability, and administrative order travel very fast indeed.

THE STORY THAT WAS BROUGHT BACK

The global propagation of the reference case cannot be explained solely by its institutional origin or administrative convenience. Those conditions enabled diffusion, but they do not explain why the framework was accepted so readily by its audiences. The decisive factor lies in the epistemic condition of those audiences themselves. The reference case arrived in environments that lacked the conceptual tools required to interrogate it. In that sense, the memeplex did not encounter resistance because there was nothing there to resist with.

At the center of this vulnerability was the near-total absence of training in representational measurement. Across economics, health economics, outcomes research, and policy analysis, the axioms governing scale types, permissible transformations, and the distinction between ordering and measuring had largely vanished from curricula by the late twentieth century. Stevens' typology was sometimes cited, rarely taught, and almost never operationalized. Rasch measurement was

regarded as niche psychometrics, not as the sole pathway to quantitative claims for latent attributes. Measurement was treated as self-evident rather than as something that had to be demonstrated.

As a result, those exposed to the NICE framework were not equipped to ask the one question that would have halted transmission immediately: *what kind of number is this?* Instead, numbers were treated as legitimate by default. If an output appeared precise, was generated by software, and could be summarized in tables and graphs, it was assumed to be quantitative. The reference case exploited this assumption perfectly. It delivered numbers with confidence intervals, sensitivity analyses, and decimal precision, all the visual markers of science, while quietly bypassing the conditions that make numbers meaningful.

What returning scholars and analysts brought back to their home institutions was therefore not a contested theory, but a narrative of modernity. The story was simple and immensely attractive. England had solved the problem of health care decision making. It had replaced political discretion with evidence. It had established a rational framework capable of comparing disparate therapies using a single metric. This was not presented as one possible approach; it was presented as *the* approach. Embedded in this story was a powerful moral claim. Decisions were no longer arbitrary. They were fair, consistent, and transparent. The QALY functioned not merely as a metric, but as a symbol of equity; equal value for equal health gain. That symbolism mattered far more than its mathematical incoherence. For audiences unfamiliar with measurement theory, the ethical narrative filled the epistemic void.

The story also promised professional legitimacy. Mastery of the reference case conferred membership in an international community of expertise. Analysts could now speak a common language, publish in recognized journals, attend ISPOR conferences, and participate in a global discourse. Acceptance of the framework became a credential. Questioning it became professionally risky. ISPOR played a decisive role at this stage. It did not invent the reference case, but it normalized it. Through Good Research Practices task forces, methodological guidelines, short courses, and journal endorsements, ISPOR transformed what had been a national administrative tool into an international professional standard. Importantly, ISPOR did not require measurement justification for this elevation. Its focus was coherence of practice, not admissibility of arithmetic. Consistency replaced validity. This institutional endorsement amplified the blind-spot. ISPOR's authority reassured audiences that foundational questions had already been addressed somewhere else. No one needed to ask whether utilities were ordinal or whether QALYs were ratio measures, because surely a society of experts would not overlook something so basic. The absence of discussion was interpreted as resolution rather than omission.

The result was a remarkable collective illusion. Each group assumed that another group possessed the missing knowledge. Academics assumed the economists had settled measurement. Economists assumed psychometricians had validated utilities. Decision-makers assumed journals had vetted the methods. Journals assumed reviewers understood the foundations. Reviewers assumed the framework was established doctrine. Responsibility dissolved into diffusion.

In this environment, the reference case was not evaluated; it was inherited. It arrived pre-legitimized, wrapped in institutional authority and reinforced by repetition. Its core claims were never subjected to falsification because falsification itself had been redefined. Instead of empirical

refutation, the framework relied on sensitivity analysis, scenario testing, and internal robustness; techniques that vary assumptions without ever challenging whether the outputs represent quantities at all.What was carried back, then, was not evidence-based medicine in any scientific sense. It was an administrative myth dressed in mathematical clothing. A story in which complexity substituted for validity, and consensus substituted for truth.

Most strikingly, the transmission succeeded because it did not demand understanding. One could apply the reference case competently without ever confronting representational measurement. Indeed, confronting it would have made application impossible. The framework's usability depended on ignorance; not malicious ignorance, but structural ignorance produced by decades of educational neglect. In this sense, the blind were not merely leading the blind. The system itself rewarded blindness. Those who questioned measurement were marginalized as philosophical, impractical, or obstructive. Those who accepted the framework were promoted as pragmatic and policy-relevant. Over time, the distinction between practicality and scientific legitimacy was erased.

The memeplex therefore propagated not because people believed something false, but because they had never been taught how to recognize falsity in the first place. When the audience lacks the concept of measurement as a precondition for arithmetic, any numerical story can pass as quantitative truth.

That is the story that was brought back. Not deception, not conspiracy, but a transferable narrative of authority without foundations. A framework that appeared to resolve uncertainty while quietly ensuring that no claim could ever be falsified. Once such a story takes hold, it does not spread through argument. It spreads through training manuals, conference slides, and methodological checklists; until, eventually, it becomes unthinkable to ask whether the numbers mean anything at all.

THE CLONING OF NICE AND THE ILLUSION OF GLOBAL CONSENSUS

One of the most unsettling findings to emerge from the Logit Working Papers is not merely that the NICE reference case spread globally, but that it did so with extraordinary uniformity. Across countries with vastly different political systems, health financing arrangements, academic traditions, and regulatory cultures, the same analytical architecture appears again and again. The same dependence on QALYs. The same use of reference-case simulation models. The same tolerance of ordinal utilities treated as interval or ratio quantities. The same absence of representational measurement as a gatekeeping requirement. The result was not diversity with local adaptation, but replication.

Each national HTA system presents itself as autonomous, context-sensitive, and locally governed. Yet when examined at the level of measurement assumptions, they are effectively clones. Their differences lie in thresholds, procedural steps, and administrative language; not in epistemic foundations. The core logic is identical. Arithmetic proceeds without prior demonstration of measurement. Modeling substitutes for falsification. Closure replaces provisional truth.

This level of convergence is difficult to reconcile with any conventional account of historical experience in scientific development. In normal science, even flawed paradigms generate dissent, schisms, rival models, and methodological debate. Here, by contrast, divergence is largely absent. The reference case did not win a contest; it appears simply to have been adopted.

The natural question follows: was the influence of NICE so overwhelming that resistance was futile? The answer is more disturbing. There was little resistance because there was little capacity for resistance. Opposition requires concepts. To resist the reference case on scientific grounds would have required widespread familiarity with representational measurement theory, scale-type admissibility, and the distinction between ordinal and quantitative structures. Those concepts were not part of the professional knowledge base of health economics, HTA, or outcomes research in most countries. They were not taught. They were not examined. They were not enforced by journals. As a result, the reference case did not confront a hostile epistemic environment; it entered a vacuum. In such conditions, adoption does not feel like surrender. It feels like alignment.

National agencies did not experience themselves as capitulating to a foreign model. They experienced themselves as joining an international best-practice community. NICE offered something enormously attractive: a ready-made analytical framework that came pre-legitimized by the UK's academic reputation, government authority, and early institutional success. For countries under pressure to make coverage decisions with limited data and finite budgets, the offer was irresistible.

What was adopted was not merely a method, but a shield. By invoking the reference case, agencies could claim that decisions were evidence-based, consistent, and internationally aligned. Responsibility shifted from judgment to procedure. Once the model had been run, the decision appeared to follow logically, even if the underlying quantities were fictional. This is why the cloning effect occurred so rapidly. The reference case solved administrative problems even as it avoided scientific ones. It provided closure without falsification. It delivered comparability without measurement. It allowed decisions to be defended procedurally even when they could not be defended empirically.

Importantly, this convergence was reinforced by the global knowledge infrastructure. Journals, academic centers, and professional societies reproduced the same assumptions because they drew from the same source materials. Students trained in one country carried the framework to another. Reviewers expected submissions to conform. Funding bodies required alignment. Over time, deviation became professionally unintelligible. This is how a memeplex achieves stability. It does not suppress alternatives; it renders them unthinkable.

From within the system, the absence of resistance appeared to confirm correctness. If everyone is doing it, surely it must be right. Consensus became mistaken for validation. Uniformity became mistaken for truth. The more countries adopted the framework, the less likely anyone was to question it. Yet what the Logit Working Papers reveal is that this apparent consensus was illusory. It was not built on shared understanding, but on shared ignorance of measurement axioms. The same foundational errors appear everywhere because the same foundational questions were never asked anywhere. This explains why the scale of acceptance was not recognized for decades. There was no instrument capable of detecting it. Traditional literature reviews cannot expose epistemic

structure. Citation counts cannot reveal conceptual absence. Peer review cannot detect what reviewers themselves do not know. As long as critique relied on human scholarship operating within the same knowledge boundaries, the memeplex remained invisible to itself.

It required a different form of interrogation.

Artificial intelligence large language models (LLM) changed the situation fundamentally. By synthesizing entire national corpora, guidelines, submissions, academic papers, training materials, LLM diagnostics made it possible to observe what had previously been unobservable: the statistical structure of belief. The probability patterns across chosen true and false canonical 24 statements revealed something no individual reader could see. Not disagreement. Not debate. But silence.

Across countries, across agencies, across journals, the same propositions collapse to the floor: measurement precedes arithmetic; multiplication requires ratio scales; latent traits require Rasch transformation. At the same time, the same false propositions rise to the ceiling: utilities as interval measures; QALYs as ratio quantities; summated scores as quantitative outcomes; simulation as evidence. The discovery is not that HTA got it wrong, which it did. It is that HTA never engaged with the question at all.

Only when belief patterns are examined at scale does the true extent of cloning become visible. What looked like plural national systems resolves into a single global knowledge structure, replicated almost perfectly from its NICE origin point. This is why the revelation is so unsettling. The world was not persuaded. It was patterned. Until now, there was no way to see the pattern. LLM resolved that issue and demonstrated not pattern similarity but the shared ignorance of the required axioms of representational measurement.

GLOBAL PROFESSIONAL FAILURE: BUT WHOSE PROBLEM?

The global diffusion of the NICE reference case and its near-universal acceptance raise an unavoidable question. If this framework violates foundational principles of measurement, why did so few academic voices challenge it? Universities are meant to be spaces of inquiry. Scholars are meant to be skeptical. Disciplines claim progress through criticism, not obedience. Yet for more than forty years, challenge was rare, fragmented, and ultimately inconsequential.

This was not a failure of evidence. The axioms of representational measurement were available. Stevens' scale typology had been published decades earlier. Rasch measurement was already well established in the human sciences. The logic that arithmetic requires measurement was not obscure, exotic, or controversial. It was foundational. The problem was not lack of knowledge in the world. It was lack of presence in the HTA knowledge base.

What emerged instead were superficial objections. Germany questioned transferability. Others debated thresholds. Some criticized utility elicitation methods. A few objected to discount rates or modeling assumptions. These critiques created the appearance of intellectual engagement while leaving the core untouched. The fundamental question "*are these numbers measures"* was never asked.

Within a memeplex, criticism is permitted only if it does not threaten the core replicators. Challenges may occur at the margins, but never at the foundation. Indeed, challenges that remain internal to the belief system serve an important function: they reinforce the system's legitimacy by demonstrating that "debate exists," while ensuring that no debate reaches the level of existential threat. In this sense, German exceptionalism to the QALY was not resistance; it was variation within captivity.

The deeper reason academics failed to intervene lies not in cowardice or incompetence, but in professional socialization. Scholars are trained within the knowledge system they inherit. Their incentives, publication, funding, promotion, relevance, depend on operating fluently within that system. Questioning its axioms carries no professional reward. Teaching them does not appear in curricula. Journals do not ask for it. Reviewers do not expect it. Students do not request it. A scholar can build an entire career optimizing within a framework without ever encountering the question of whether the framework itself is coherent.

This is precisely how memeplexes persist. Memeplexes do not survive by suppressing dissent through force. They survive by shaping what counts as a sensible question. They establish cognitive boundaries that make certain thoughts appear unnecessary, irrelevant, or even unintelligible. Within such a system, asking whether utilities are ordinal rather than interval does not sound revolutionary; it sounds pedantic. Asking whether QALYs can be multiplied does not sound profound; it sounds obstructive. Asking whether simulation outputs can be falsified does not sound scientific; it sounds impractical. The memeplex reframes foundational critique as bad behavior.

This is why curiosity failed. Not because academics stopped being curious, but because curiosity itself was redirected. Scholars became curious about better algorithms, more refined preferences, more nuanced heterogeneity, and more sophisticated uncertainty analysis ; all downstream of an assumption that was never itself permitted to be questioned. They explored endlessly *within* the system, while never stepping *outside* it.

In retrospect, the failure appears astonishing. Yet it is entirely consistent with how belief systems behave once institutionalized. Once HTA became embedded in government processes, professional societies, degree programs, and journals, it ceased to function as a provisional scientific hypothesis and became an administrative technology. At that point, challenging it no longer appeared as scientific inquiry, but as destabilization; not an activity bureaucratic science will reward.

The emergence of AI-based LLM diagnostics marks a rupture precisely because it operates outside the social incentives that sustained the memeplex. It does not need to publish. It does not need funding. It does not fear rejection. It simply reads everything and reports what is there and what is not. In the HTA memeplex it reveals is not disagreement, but silence.

CONCLUSIONS

Across countries, journals, agencies, and academic centers, the same silences recur with statistical regularity. Measurement precedes arithmetic is not debated; it is missing. Rasch is not rejected; it

is invisible. Representational measurement theory is not controversial; it is absent. This is not a series of individual oversights. It is a collective epistemic failure.

But it is also an opportunity. Because once the structure of belief is exposed, it cannot be unseen. The emperor does not merely lack clothes; the population now has a measurement instrument capable of demonstrating that absence reproducibly, across institutions, across nations, across time and between emperors.

The question is no longer whether the reference case is defensible. It is whether the profession is willing to acknowledge that it has confused administrative convenience with scientific legitimacy for four decades. That failure belongs to no single agency, journal, or country. It belongs to the memeplex that allowed arithmetic without measurement to become normal, and curiosity without foundations to become acceptable.

The task now is not to assign blame, but to restore the conditions under which science can function again: explicit axioms, admissible measures, falsifiable claims, and the humility to accept that even widely believed systems can be wrong. HTA now has no other future.

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] Popper K. *Objective Knowledge: An Evolutionary Approach*. Revised edition. Oxford: Clarendon Press, 1979.

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

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

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