MAIMON WORKING PAPER No 14 August 2025

NUMERICAL STORYTELLING: THE PSEUDOSCIENCE OF PBAC COST-EFFECTIVENESS CLAIMS

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

ABSTRACT

This paper presents a systematic critique of the Pharmaceutical Benefits Advisory Committee (PBAC) and its long-standing endorsement of reference case modeling as the foundation for cost-effectiveness claims in health technology assessment (HTA). It argues that PBAC methodology is not grounded in science but in numerical storytelling: the simulation of imaginary futures built from invalid preference scores, arbitrary assumptions, and non-replicable models

The foundations of representational measurement theory (RMT), grounded in formal axioms established by Krantz, Luce, Suppes, and Tversky (1971), remain ignored in health technology assessment (HTA), relegating the field to pseudoscience. RMT defines measurement as the mapping of empirical relations to numerical systems through homomorphisms, emphasizing that numbers represent qualitative relations rather than intrinsic properties. Stevens's typology of nominal, ordinal, interval, and ratio scales corresponds to different permissible mappings, yet HTA persists in treating ordinal utility scores, such as those derived through time trade-off (TTO) methods, as if they were interval or ratio scales.

This misapplication crystallizes in the QALY: multiplying ordinal preference scores by ratio-scale time produces a mathematical fiction, not a quantifiable measure. HTA agencies, including the PBAC, NICE, ICER, and CMA, embrace "numerical storytelling," building cost-effectiveness models reliant on untestable assumptions, simulated futures, and internally consistent yet epistemically vacuous outputs. Probabilistic sensitivity analysis merely masks ignorance with an illusion of precision.

All meaningful scientific measurement must adhere to unidimensionality, linearity, and invariance, supporting arithmetic operations. Only interval or ratio scales or, for latent constructs, Rasch-transformed logit scales, meet these requirements. The Rasch model uniquely converts ordinal responses into linear, invariant measures suitable for comparison and arithmetic. Yet HTA continues to promote multi-attribute utility instruments like EQ-5D-3L, ignoring these fundamental measurement imperatives.

HTA's reliance on the reference case model, QALYs, and the resulting imaginary and impossible cost-effectiveness claims, constitutes a belief system, a meme, sustained through institutional inertia rather than epistemological legitimacy. Its outputs are not evidence but simulation, lacking falsifiability, replicability, or empirical grounding. To evolve into a true empirical science, HTA must reject composite preference scores and speculative modeling. Instead, value claims must be based on valid measurements: direct ratio scales for observable phenomena and Rasch-based interval measures for latent constructs.

Only then can HTA shift from numerical fiction to objective knowledge, grounded in empirical validation, transparent protocols, and continuous evaluation. The future of HTA depends on abandoning the illusion of measurement and embracing the rigorous discipline of science.

INTRODUCTION: REPRESENTATIONAL MEASUREMENT

In the practice of health technology assessment (HTA) the axioms of fundamental measurement are of no apparent interest; a position maintained for over 40 years. This puts HTA in a unique pseudoscience category. While these axioms were formalized and accepted by the 1970s in wat is described as representational measurement theory (RMT), they are of no interest as far as HTA is concerned [1]. The importance of these axioms cannot be overstated: RMT formulates the rules of quantifications despite.

RMT makes clear that measurement is the assignment of numbers to empirical entities based on their structural relationships, rather than inherent numerical properties. This approach involves constructing homomorphisms, structure-preserving mappings, from empirical relational structures to numerical systems. For instance, if objects can be ordered by weight, and combining weights is meaningful, then a numerical representation can be established where addition reflects this combination. RMT emphasizes that numbers serve as representations of qualitative relations, not as intrinsic attributes. This theory has been influential in formalizing measurement across various scientific domains, providing a framework for understanding how numerical scales correspond to empirical observations. While RMT has faced criticism for its abstract nature and limited applicability to practical measurement scenarios, especially where uncertainty and error are prevalent, RMT remains a foundational concept in the philosophy of measurement, highlighting the importance of structural correspondence between empirical phenomena and their numerical representations

The representational theory of measurement defines measurement as a structure-preserving mapping, an isomorphism, between empirical relations and numerical systems. Stevens's 1946 typology, nominal, ordinal, interval, and ratio scale, aligns neatly with this framework because each scale type corresponds to a specific class of mappings that preserve certain empirical relations and permit specific numerical operations. In representational terms, nominal scales preserve only class membership, ordinal scales preserve ordering, interval scales preserve equal intervals, and ratio scales also preserve an absolute zero. Thus, Stevens's levels can be seen as concrete instantiations of the representational theory's abstract notions of permissible transformations and structural correspondence

Unfortunately, despite the acceptance of Stevens typology in setting boundaries for measurement types and the insights and rules established by RMT for empirical claims assessment, these efforts over some 70 years passed by unnoticed by HTA. Instead, HTA insisted on a path that guaranteed measurement failure: the valuation of health state descriptions through techniques such as time trade-off (TTO).

This absurd decision sets the stage for the inevitable negative critiques of reference case models and the various guidelines put in place by gatekeepers such as the Pharmaceutical Benefits Advisory Committee (PBAC). While it easy to dismiss these endeavors out of hand, it is important

to understand why they are nothing more than numerical storytelling. This critique proceeds not as a difference of opinion but as a necessary demolition. The PBAC's framework fails on three core scientific grounds: measurement, epistemology, and methodological legitimacy. It is not science; it is spreadsheet theater that has a global audience and cheerleaders [2] . This is the purpose of this working paper.

THE BANE OF NUMERICAL STORYTELLING

The PBAC is not alone in its embrace of numerical storytelling. Across the global health technology assessment (HTA) landscape, national and regional agencies have institutionalized reference case modeling as the methodological standard for evaluating cost-effectiveness. Yet beneath the formalized spreadsheets and probabilistic analyses lies a profound departure from science. What these agencies endorse is not measurement, but narrative construction, imaginary futures assembled through chains of untestable assumptions. This practice deserves its proper name: numerical storytelling.

Numerical storytelling refers to the use of simulation models to fabricate outcomes from assumptions and preference scores that lack valid measurement properties. These models construct counterfactual futures based on what are called "believable assumptions" a phrase that signals methodological failure. Believability is not science. It is a rhetorical shield for ignorance, confusion, and convenience. The models then report cost-effectiveness ratios and quality-adjusted life years (QALYs) as though they were empirical observations. They are not. They are artifacts of invention: products of simulation, not discovery.

The fact that a model is internally consistent does not mean it maps onto reality. Yet the PBAC, and its counterparts such as NICE, ICER, and CADTH, proceed as if this epistemic chasm does not exist. It is a blindness born of institutionalized comfort: a regime where assumptions replace evidence and where ignorance of measurement theory by practitioners and their audience ensures no one notices the fraud.

The origins of this failure are not recent. When the first PBAC guidelines were issued over 30 years ago, the axioms of representational measurement were well established. The limitations of ordinal scales, and the impossibility of multiplying them with ratio scales like time, were not obscure truths; they were, and remain, elementary. That PBAC ignored these axioms then, and continues to do so today, reflects not oversight but systemic intellectual failure. This is not the first time the question of the standards of normal science have been proposed as the standard the PBAC should endorse. In 2017 a critique was presented of the just published and still current Version 5 of the PBAC guidelines pointing out that they failed the required standards for claims to be credible evaluable and replicable [3]. The critique made clear that constructed claims for product impact were unacceptable. Although the PBAC were aware of this critique, nothing happened.

THE IMPERATIVE OF MEASUREMENT IN SCIENCE

Science begins with observation, but it only becomes science through measurement. Without a standard, rules by which to quantify what is observed, there can be no testable claim, no replication, and no comparison. Measurement is not a peripheral consideration; it is the foundation on which

all empirical knowledge rests [4]. And with that foundation come strict, unyielding rules. These rules are not flexible. They are axiomatic.

The laws of measurement theory divide data into scale types: nominal, ordinal, interval, and ratio. Only the last two support meaningful arithmetic operations. Interval scales permit the assessment of equal differences; ratio scales permit comparison of absolute magnitudes, anchored by a true zero. These are not arbitrary distinctions. They define what it means to quantify. A value claim that does not rest on at least an interval scale is not a claim about quantity, it is a gesture, an opinion dressed in numbers.

It is, however not just categorizing data series. We have to go deeper and establish rules for quantification; hence the importance of RMT; we have to assign numbers to empirical entities based on their structural relationships. This is the first and most egregious failure of PBAC-style modeling: the routine use of ordinal utility scores, such as EQ-5D or SF-6D, as if they were interval or ratio measures. The resulting QALY, constructed by multiplying these ordinal scores by time (a ratio scale), is a mathematical fiction. The product is not a quantity; it is an error. No amount of face validity, expert consensus, or probabilistic sensitivity analysis can repair this mistake. You cannot rescue bad mathematics with better spreadsheets.

There are no exceptions to this imperative. All measures that purport to represent reality, whether in physics, biology, psychology, or health economics, must meet this requirement. Observable constructs must be measured on linear, unidimensional, interval or ratio scales. Latent constructs, those that reflect subjective or internal states, must be transformed into measures using models that respect the same imperatives.

Here the Rasch model stands alone [5]. It offers a rigorous framework of axioms or rules to convert ordinal responses into linear, invariant measures [6]. Rasch measurement is not a psychometric convenience; it is the scientific equivalent of a physical instrument for latent variables. By anchoring item difficulty and respondent ability on a common logit scale, the Rasch model constructs an interval framework for the subjective; achieving constant relative differences that permit valid comparison. It is the best example of the application, in probabilistic and not deterministic terms, of the rules of quantification to establish a measurement structure for latent traits. A model that was developed essentially independently of the mainstream focus on what was later termed RMT.

Whether one is quantifying force or fatigue, speed or symptom severity, the rule is the same: measure or fail to know. Claims made without valid measurement are not scientific claims. They are conjectures; untestable, non-replicable, and epistemically bankrupt. That is the standard PBAC has abandoned. And that abandonment is not a matter of method, it is a matter of science. In the sections that follow, we examine what happens when that standard is ignored: the collapse of epistemology, the rise of simulation fiction, and the loss of credibility in health technology assessment.

THE PBAC STORY: THE HAPPY ENDING FOR NUMERICAL STORYTELLING

The PBAC guidelines prescribe a constrained modeling paradigm centered on economic simulation, embedding two primary structures: cohort-based state transition models and individual-level microsimulation [7] These are positioned as default tools for projecting long-term cost-effectiveness, with selection criteria driven not by measurement theory but by the complexity of disease progression and computational tractability.

Cohort-based state transition models, typically Markov structures, are to be employed where disease trajectories can be partitioned into a finite, manageable number of mutually exclusive health states. Transitions between states must conform to the Markovian "memoryless" assumption unless justified via tunnel states that mimic time-dependent or history-dependent behaviors. Submissions must specify the rationale for transition probability assumptions, justify cycle length selection, and apply a half-cycle correction unless an alternative is defensible. Where patient heterogeneity violates modeling assumptions (e.g., non-normal distributions of age), stratified analyses are mandated. The analytical logic is that these corrections will preserve the predictive validity of the model; however, this belief rests on untestable and unmeasurable assumptions.

Microsimulation or discrete event simulation is permitted only when the disease or intervention complexity exceeds the representational capacity of a cohort model. Factors warranting such an approach include time-varying hazard rates, patient histories influencing future risk, or continuous disease markers. Yet, even here, the requirement is to justify structural complexity within the same overarching cost-utility framework.

Layered atop this structural modeling is the mandatory presentation of outcomes in terms of a final health metric to serve as the denominator in the base-case incremental cost-effectiveness ratio (ICER). This is almost invariably the quality-adjusted life year (QALY), derived from multiattribute utility instruments (MAUIs) or indirectly via mapping algorithms. PBAC requires the use of Australian-based preference weights where possible and mandates full transparency regarding the derivation and transformation of patient-reported outcomes into utility values.

The guidelines insist that both directly elicited and mapped QALYs be presented and compared, even where direct measurement is methodologically flawed or where mapping is empirically unjustifiable. Submissions must report point estimates, standard deviations, and confidence intervals for utility weights, reinforcing the illusion of precision in a model architecture that is inherently unverifiable. Non-patient health outcomes, such as caregiver burden or quality-of-life impacts on family, are explicitly excluded from the base case, signaling a rigid boundary around what qualifies as "economic" benefit.

The reference model results should present the estimated incremental cost, incremental outcome(s), and the resulting cost-effectiveness ratio(s), typically expressed as cost per QALY gained. Where applicable, both cost-utility analysis (CUA) and cost-effectiveness analysis (CEA) results should be included. If the ICER is based on an outcome other than QALYs or life-years (e.g., hospital days avoided or cases prevented), sponsors should compare these results with

previous PBAC decisions that used the same or similar outcome measures to support consistency and interpretability across submissions.

To understand the pathology of numerical storytelling, we must first understand its purpose. Within the PBAC framework, the model is not just a technical artifact; it is the very engine of decision-making. The model *tells a story*, and the story must lead to a happy ending: a product deemed acceptably cost-effective, within a predefined threshold of willingness-to-pay per QALY gained. This is the ritualized logic of PBAC evaluation.

The PBAC does not ask for evidence in the classical scientific sense, observable, replicable, falsifiable data from the world. Instead, it asks for *modelled evidence;* a simulated narrative of future health states, costs, and quality-of-life values. The model is constructed around a core question: "*Given these assumptions, what would the long-run cost-effectiveness ratio be, if this imaginary world were true*?" This is not a scientific question. It is a speculative fiction. The resulting model is not a report of what has happened or what can be known; it is a constructed reality populated by guesses, preferences, and extrapolations.

This is not hypothesis testing. This is narrative engineering. The model is not a representation of reality, it is a tool to produce a politically palatable output: a number, finely tuned to suggest value. The assumptions driving the story are rarely questioned. They are granted the status of "believable inputs" so long as they are judged to be internally coherent and superficially justified. There is no formal requirement for empirical evaluability, no test of whether these assumptions will ever reflect observed outcomes in real populations. No demand for replication. No standard for falsification. A rejection not only of the standards for normal science by a relativist number game that ignores the theory of measurement.

Moreover, the PBAC model encourages sponsors to obscure uncertainty through the ritual of probabilistic sensitivity analysis (PSA). By running thousands of simulations on uncertain parameters, the sponsor creates a cloud of outcomes, typically presented as a cost-effectiveness acceptability curve. This curve does not clarify anything; it simply quantifies ignorance. It gives the illusion of confidence, while burying the absence of empirical grounding in a sea of statistical output. This is not measurement. It is theatre.

In sum, the PBAC story is structured not to discover provisional claims and meaningful pricing, but to *arrive at a decision*. The goal is not empirical accuracy, but administrative acceptability. The model must produce a result that appears to justify reimbursement within a pre-established budgetary narrative. And so long as the story ends with an ICER below the threshold, the structure of that story, no matter how incoherent or unmeasurable, escapes scrutiny.

This is the essential danger: decisions are being made, policies enacted, and resources allocated based on simulated futures with no claim to epistemic legitimacy. The PBAC does not merely fail to demand science. It incentivizes its opposite: the construction of agreeable fictions. And those fictions are only possible because the PBAC has never enforced or been asked to enforce the scientific imperative of valid measurement.

In the end, what masquerades as a cost-effectiveness model is best understood as a marketing exercise. It is a submission crafted not to test a hypothesis, but to persuade a committee. The narrative is built backwards from a desired conclusion, cost-effectiveness below a threshold, and the parameters are tuned to achieve that end. This is not analysis in the service of truth; it is storytelling in the service of access. And the PBAC, rather than policing this behavior, has made it the default.

DECONSTRUCTING THE EQ-5D-3L PREFERENCE SCORE

At the heart of the PBAC's cost-effectiveness framework, and indeed, at the heart of global HTA practices, lies a singular fiction: the EQ-5D-3L preference score [8]. This score is used to generate quality-adjusted life years, the central outcome in most reference case evaluations. It purports to measure "health-related quality of life" and is presented as a continuous scale anchored at zero for death and one for perfect health. But the truth is far more damning: the EQ-5D-3L preference algorithm fails even the most basic standards of measurement theory. It should never have been used. It should have been abandoned at its inception on a Greek hillside above the snow line.

The EQ-5D-3L consists of five items covering different aspects of health—mobility, self-care, usual activities, pain/discomfort, and anxiety/depression—each scored at three ordinal levels. These items are treated as if they can be aggregated into a composite representation of health status, but no attempt is made to demonstrate unidimensionality. The items span distinct health domains without evidence that they reflect a single latent construct. The response levels within each domain are ordinal and unevenly spaced, with no justification that differences between levels are consistent either within or across domains. These responses, fundamentally ordinal, are then transformed through an arbitrary weighting system based on societal preferences.

The preference weights themselves are derived using the time trade-off (TTO) method, where respondents are asked to imagine trading off years of life to avoid living in various impaired health states. This method introduces a further abstraction, relying on imagined choices in hypothetical contexts. The responses are unstable, context-dependent, and deeply influenced by framing effects. More critically, TTO responses do not, and cannot, produce ratio or interval scale outputs. They are ordinal composite rankings at best. When averaged and forced through regression models to create a utility function, the result is an algorithm that yield only numbers. These have no reference point in measurement theory; they are mathematically incoherent. The utility or preference score lacks a true zero, lacks meaningful units, and has no evidence of invariance across populations or contexts.

This entire process violates every axiom of fundamental measurement. There is no demonstration of linearity. There is no evidence of constant relative differences. There is no claim to invariance. And there is no attempt, nor any theoretical basis, for asserting ratio properties. The EQ-5D-3L is not a measure; it is an opinion artifact. It masquerades as a quantification of health but is, in fact, a sequence of ordinal responses transformed into a numerical fiction by processes wholly divorced from the science of measurement.

The criteria for any valid measure are clear: unidimensionality, linearity, invariance, and the ability to support arithmetic operations. The EQ-5D-3L, with its associated family of instruments, meets

none of these. It does not even reach first base. From the moment it was proposed, it should have been rejected on methodological grounds alone. But instead of being disqualified, it was institutionalized, and with it, the entire structure of the QALY and the reference case model became anchored in a pseudo-measure that has no scientific legitimacy.

NONSENSE ON STILTS: THE REFERENCE CASE DECISION SIMULATION

The PBAC reference case does not represent an evolution of scientific reasoning in health technology assessment; it is a regression to pre-scientific belief systems. It constructs decisions from speculative models, unmeasurable inputs, and invented outcomes. It rests not on observation, not on empirical testing, not even on coherent arithmetic, but on simulation. Worse still, it treats this simulation as if it were evidence. This is not a matter of technical refinement. It is a wholesale abandonment of the standards of science.

At the core of the PBAC reference case is a model-based framework that simulates costs and outcomes over a hypothetical patient lifetime. This framework is designed to accommodate gaps (i.e., fill in non-existent data) at the time of product launch. Rather than calling for investment in empirical observation or the development of a research program to generate meaningful data, the reference case invites immediate and untested speculation. The sponsor is not asked to provide evaluable evidence but to simulate it. The justification for this system is practical expediency: limited data must be overcome with "best available evidence" and "reasonable assumptions." In reality, it is a mechanism for fabricating imaginary claims while maintaining the outward appearance of rigor.

There is no concept of prediction in the reference case that is subject to falsification. The entire structure is designed to prevent that possibility. Claims are constructed about events in a counterfactual future and assessed not by their truth but by their plausibility. These modeled outputs are treated as if they were empirical findings, despite being produced by processes that cannot be replicated outside the simulation itself. Assumptions are substituted for data, and internal coherence is mistaken for external validity. This would be unacceptable in any other domain of science. It should have been rejected out of hand, just as the EQ-5D-3L should have been.

The failure of the reference case is dual: it violates both the axioms of fundamental measurement and the epistemological standards of normal science. Measurement theory tells us that only interval and ratio scales can support arithmetic operations, including multiplication and division. Science tells us that knowledge claims must be credible, evaluable, and replicable. The PBAC reference case fails on both counts. It uses inputs that are not legitimate measures, such as preference scores derived from ordinal data, and it generates outputs that cannot be tested or reproduced in the real world. It is as though the scientific revolution of the seventeenth century had never occurred [9]. The principles laid down by Galileo, Descartes, Newton, and later formalized by Popper, the need for testable hypotheses, falsifiability, empirical replication, have been cast aside. In their place we find a parody of science: so-called approximate information masquerading as decision support.

The linchpin of this simulation framework is the QALY. But the QALY is impossible. It is mathematically incoherent from the moment it was conceived. It rests on the assumption that health states can be assigned values on an interval or ratio scale and then multiplied by time. Yet

the so-called values in the QALY are not values at all, they are ordinal preferences, often derived through time trade-off (TTO) exercises. These exercises ask individuals to rank hypothetical health states or trade length of life for quality. The resulting scores are averaged and transformed, often through linear regression, into a single-number utility. But these scores have no demonstrated scale properties. They do not have equal intervals. They do not support ratio comparisons. They cannot be meaningfully multiplied by time, which *is* a ratio scale. This is not an oversight; it is a categorical error.

The deeper error is conceptual. The QALY pretends to quantify a latent construct, overall health-related quality of life, by aggregating multiple domains of function and feeding into a single score. But a latent construct cannot be multidimensional. It cannot be built from conceptually distinct components, this fails the requirement for dimensional homogeneity where time, a ratio scale, cannot be combined with an ordinal discount factor. If a latent trait is to be measured, it must be unidimensional. It must represent a single underlying continuum, revealed through a series of related observations. It must satisfy the axioms of fundamental measurement. And if the responses are ordinal, as they always are in subjective assessments, then they must be transformed through a model that yields a linear, logit interval scale.

Only one framework meets that criterion: the Rasch model. The Rasch model transforms ordinal responses into linear logits, placing both item difficulty and person ability on a common ratio scale. It guarantees unidimensionality through fit statistics and enforces constant relative differences across the trait continuum. It is the only way to construct a measure from subjective data. Yet the QALY rejects this framework entirely. It bypasses measurement and substitutes preference. It assumes that the expression of preferences can serve as a surrogate for measurement, even when the mathematical properties of the resulting score are unknown or undefined.

Even if the EQ-5D-3L had met Rasch standards, and it does not, it would still have failed. For while Rasch provides a linear *logit* scale, one cannot multiply a logit by time. The logit represents the logarithm of the odds of success on the latent trait continuum; it is a relative scale, not an absolute one. Multiplying time by a logit is as meaningless as multiplying time by temperature in Celsius or decibels of sound pressure. The QALY is thus not just invalid, it is mathematically impossible. It is the product of a misunderstanding so deep that it calls into question the scientific legitimacy of every model that incorporates it; which amounts to literally tens of thousands of peer reviews publications over the past 35 years.

The reference case decision simulation is built upon this impossibility. It converts non-measures into metrics, aggregates unquantifiable components, and projects them into an imagined future. The output is treated as a decision rule, a signal of value, a justification for access, a summary of therapeutic worth. But it is none of these. It is numerical storytelling. It is nonsense on stilts.

We now ask what must follow from this collapse, what a scientifically legitimate alternative would look like, grounded in protocols, real-world data, and the principles of fundamental measurement?

A NEW START IN HTA: FROM MEME TO PARADIGM

The time has come to confront an uncomfortable truth: the framework that the PBAC, and its global counterparts, have embraced for over three decades is not a scientific paradigm. It is a meme. It replicates not because it is true, but because it is familiar, convenient, and institutionally reinforced. It is a belief system, not a body of knowledge; one sustained by rhetoric, persuasion, and authority. It reflects a relativist, post-modern stance in which science is no longer seen as the path to understanding reality, but as one narrative among many. In this belief system, evidence is not discovered through observation; it is invented through simulation.

It is a belief system, enforced and perpetuated by professional associations such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), which ensure unwavering fidelity to the core doctrines of numerical storytelling across successive generations of students and their instructors. Central to this indoctrination is the promotion of imaginary guidelines, most notably the CHEERS 2022 encyclical, which offers detailed instructions, Markov model-driven, QALY-based, on how to construct fictional reference case model narratives [10]. A checklist is even provided to standardize the fabrication. Leading journals, far from challenging this pseudoscience, have eagerly endorsed it, turning publication into a ritualistic validation of imaginary constructs rather than a test of scientific claims. The entire enterprise exists not to advance evaluable knowledge but to sustain a self-reinforcing cycle of fabricated authority.

The reference case model, the QALY, and the ICER are not the products of disciplined scientific inquiry. They are artifacts of a methodological regime that has confused internal logical consistency with external validity, and narrative coherence with empirical credibility. It is a textbook example of what sociologists of science call a “strong program”; a system that survives through institutional inertia, social consensus, and the suppression of dissent, not through alignment with truth [11]. It is time to end the pretense that this is science; the belief that evidence is never discovered but constructed within a particular social community, whether an academic community or a Rastafarian commune [8]. It is time to dismantle the meme and restore the imperative of objective measurement and the appeal to superior evidence.

What has evolved under the banner of health technology assessment is not a method of discovery, but a method of decision simulation based on the mathematically incoherent valuation of health state descriptions. It survives through repetition, policy inertia, and the illusion of precision. Its defenders have never understood the axioms of representational measurement theory. They are unaware. or deliberately dismissive, of the problem of induction. They believe, wrongly, that a model that conforms to its own assumptions is therefore informative. They speak of “best available evidence” while ignoring that what they are producing is not evidence at all. It is numbers masquerading as measurement, speculation disguised as knowledge.

This is not how science works. A true paradigm is constrained by rules; by logical structure, by empirical accountability, by the imperatives of replication and falsification. It does not rest on assumptions it cannot test, or rely on scales it cannot defend. And it certainly does not rely on constructs, such as the QALY, that cannot meet even the most basic requirements of arithmetic. The pretense that such models represent "value" is indefensible. They cannot, because they do not measure anything. The underlying preference scores are not linear. The utilities are not invariant. The entire system is a performance, not a science.

The future of HTA depends on recognizing this foundational failure. It requires abandoning the meme and constructing a legitimate paradigm; one grounded in the axioms of representational measurement and the empirical discipline of science. Unlike a meme, a paradigm is a structured framework of inquiry, built on testable hypotheses, valid measurement, and replicable evidence. It imposes constraints. It demands adherence to the principles of empirical investigation, the rigor of observation, and the standard of falsifiability. A paradigm does not survive by repetition or rhetorical appeal; it endures by producing objective knowledge. Where a meme spreads through imitation, a paradigm is sustained through falsification. It is the architecture of science defining not only what counts as evidence, but how claims are constructed, tested, and judged.

From the perspective of HTA, such a paradigm would focus on the development of value claims that are evaluable, replicable, and aligned with the rules of measurement [12]. It would reject composite metrics, preference-based scoring systems, and simulation models that invent outcomes rather than discover them. Instead, it would demand value claims founded on valid interval or ratio measures, either through the direct observation of resource utilization or through Rasch-transformed instruments that yield linear, invariant measures of latent traits. This paradigm would prioritize transparency, prospective protocol design, and ongoing review through therapeutic class and disease area assessments. HTA would be reconstituted as an empirical science, no longer a ritual of policy justification, but a discipline accountable to evidence, measurement, and replicable inquiry.

The new paradigm would begin by recognizing that in HTA there are only two defensible forms of measurement. The first is the linear ratio scale, used to quantify directly observable phenomena such as time, units of care, or behavioral compliance. The second is the Rasch logit ratio scale, which alone can transform ordinal responses into valid interval measures when evaluating latent constructs such as need fulfillment, treatment satisfaction, or functional burden. There are no alternatives. The science of HTA must begin here; or not at all.

There are no other options. There are no middle paths, no alternative models that evade these constraints. Either a value claim meets the axioms of measurement and the criteria of science, or it does not. If it does not, then it is not a value claim; it is an artifact of convenience, built to satisfy administrative rituals rather than scientific standards. A value claim, properly defined, is a proposition about the expected impact of a therapy, clinical, behavioral, or patient-reported, that can be expressed in measurable terms, subjected to empirical evaluation, and reproduced under a defined protocol. It must be founded on data that conform to the rules of measurement: linear, unidimensional, and ratio-scaled if observed directly, or transformed to Rasch logit ratio form if based on ordinal reports of subjective experience. A value claim is not a simulation of a future scenario. It is a testable assertion about a therapy's effects in a defined target population, evaluated within a specified timeframe, using data collected under transparent and replicable conditions. Its legitimacy rests not on whether it can be modeled, but whether it can be measured. That is the defining line between science and numerical storytelling, between meaningful evidence and manufactured belief.

This transition from meme to paradigm is not merely academic. It is a structural imperative. As long as HTA institutions continue to endorse models that cannot produce evaluable claims, they will remain agents of misinformation, complicit in the allocation of billions in public funds on

the basis of numerically grounded fiction [13]. The credibility of the field depends not on refining the simulation but on rejecting it.

CONCLUSION: THE EVOLUTION OF OBJECTIVE KNOWLEDGE

If the PBAC, or any future gatekeeper HTA agency in Australia that aspires to a foundation in science has to come in from the cold of pseudoscience, it must abandon the comfortable fictions that have sustained HTA for decades. The reference case, the QALY, the ICER, and the entire architecture of numerical storytelling have no place in a science of HTA. If HTA is to mature into a discipline of objective knowledge rather than remain a system of administrative justification, it must sever itself from these artifacts of convenience [14]. The transition will not be easy. Memes do not collapse quietly. But collapse they must if progress is to occur. The evolution of objective knowledge demands more. It demands rigor. It demands fidelity to measurement theory and epistemological discipline. And it demands an institutional architecture capable of saying no to fiction, even when that fiction is elegant, convenient, or politically expedient. There is no place for cost-effectiveness models that rely on ordinal preference scores, no matter how elaborate their sensitivity analyses or how authoritative their presentation.

The heart of this transformation is the rejection of multiattribute instruments and reference case modeling as supposedly legitimate scientific methodologies. The claim that health can be summarized by a single index, constructed from ordinal preferences and then manipulated through simulation, is not just flawed; it is an anachronism. It reflects a pre-scientific conception of knowledge, one that confuses arithmetic for measurement and coherence for truth. If we do not escape this structure, we are condemned to an eternity of CHEERS circular reasoning: continually rerunning lifetime simulations, inventing plausible future pathways, and assigning prices to imagined states of being. It is the intellectual equivalent of watching *The Pirates of Penzance* operetta on an endless loop, entertaining perhaps to Gilbert and Sullivan disciples, but entirely detached from reality.

To exit this Gilbertian loop, the PBAC, or any successor agency, must adopt a new orientation: a commitment to what Popper described as the evolution of objective knowledge. Objective knowledge is not personal belief or institutional consensus; it is knowledge that exists independently of its creators, tested and refined through critical scrutiny, empirical challenge, and public replication. That orientation must begin with a commitment to addressing evidence gaps, not papering over them with inductivist models. Where data are lacking, the proper response is not to invent assumptions drawn from the literature, but to invest in evidence generation. This means developing protocols for real-world evaluations, establishing benchmarks, and accepting that value claims must be provisional; subject to revision as new data emerge. A claim that cannot be falsified or replicated is not a value claim; it is a hypothesis posing as a conclusion.

This will require far more than a cultural shift within agencies like the PBAC. It demands a *Weltanschauung* transformation, a fundamental rupture with the institutional worldview that has treated modeling as evidence and simulation as science. There is no single solution or universal model that can deliver final truth. The pursuit of objective knowledge in Australian HTA must be a continuous process, grounded in the design and execution of evaluable claims, tested in real populations, and subject to ongoing therapeutic class reviews and disease-area monitoring. Claims

must be credible, evidence-based, and replicable. They must rely on appropriate scales: ratio measures for directly observed behaviors such as compliance or resource utilization, and Rasch logit ratio scales for subjective constructs such as patient-reported outcomes. There are no shortcuts. Measurement, not modeling, is the foundation of knowledge.

Inevitably, there will be pushback. Disciples are not easily re-educated**.** After more than three decades of relativism embedded in the reference case, the simulation model has become not merely a methodological choice, but the axis of professional identity, ideological comfort, and institutional legitimacy. Its disciples will not relinquish it quietly. They will assert its pragmatism, appeal to the absence of perfect data, and argue that simulation is the best available option. But this defense no longer holds. It is not that the models are insufficient; they are scientifically wrong. They produce outputs that are not measurements, predictions that cannot be tested, and decisions that rest on foundations of numerical fiction. Their continued use does not merely perpetuate error; it compromises the very legitimacy of HTA as a scientific discipline**.**

This is not a rejection of progress. It is its very condition. To advance, we must strip away the assumptions that have shielded us from scrutiny and commit to a science that is falsifiable, testable, and real. This means recognizing that HTA is not about one-off decisions at the moment of launch, but about building a platform for continual assessment; a commitment to measuring therapy impact over time, across populations, and within evolving therapeutic landscapes.

The choice is now before us. We can cling to the meme, rehearse the ritual of conferences and restaurant meetings, and maintain the illusion. Or we can pursue the harder, slower, but ultimately more rewarding path of science. The PBAC must choose. It can remain a steward of imaginary futures or it can evolve into a paradigm for objective knowledge. The moment for that choice is now.

ACKNOWLEDGEMENT

The development of this paper has benefited from the use of AI-assisted tools, specifically ChatGPT by OpenAI (https://chat.openai.com/), Version 5, for tasks including revised drafting and text editing. The author takes full responsibility for the content and any errors that may remain

REFERENCES

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

[2] Drummond M, Sculpher M, Claxton K, et al. Methods for the Economic Evaluation of Health Care Programmes (4th Ed). New York: Oxford University Press, 2015

[3] Langley P. Dreamtime: Version 5.0 of the Australian Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee (PBAC). *Inov Pharm*. 2017; 8(1): Article 5

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

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

[6] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research. 1960

[7] Australian Government. Department of Health and Aged Care. The Pharmaceutical Benefits Advisory Committee Guidelines. Version 5.0, September 2016

[8] EUROQOL https://euroqol.org/information-and-support/documentation/user-guides/

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

[10] Husereau D, Drummond M, Augustovski F, *et al.*: Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) Statement: Updated reporting guidance for health economic evaluations. *ValueHealth.* 2022; **25**(1): 3–9.

[11] Bloor D. Knowledge and Social Imagery (2nd Ed). Chicago: University of Chicago Press, 1991

[12] Langley P. Nothing to Cheer About: Endorsing Imaginary Economic Evaluations and Value Claims with CHEERS 22 [version 1; peer review: 2 approved]. *F1000Research* 2022, **11**:248 (https://doi.org/10.12688/f1000research.109389.1)

[13] Langley P. Facilitating bias in cost-effectiveness analysis: CHEERS 2022 and the creation of assumption-driven imaginary value claims in health technology assessment [version 1; peer review: 3 approved]. *F1000Research* 2022, **11**:993 (https://doi.org/10.12688/f1000research.123709.1)

[14] Popper K. Objective Knowledge: An Evolutionary Approach (Rev Ed). Oxford: Clarendon Press, 1979