MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED KINGDOM: THE DEATH OF MULTIATTRIBUTE INSTRUMENTS IN HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 202 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

Multiattribute utility instruments such as the Health Utilities Index, EQ-5D, and the Australian Assessment of Quality of Life (AQoL) have long served as the quantitative foundation of health technology assessment (HTA), supplying preference-based utilities for the construction of quality-adjusted life years and cost-effectiveness claims. This paper argues that these instruments have no scientific role in HTA. The conclusion is not normative or policy-driven; it follows directly from the axioms of representational measurement and is confirmed empirically through canonical logit interrogation of instruments, journals, and national HTA frameworks.

Using a fixed 24-item diagnostic grounded in unidimensionality, scale-type admissibility, invariance, Rasch measurement, and falsifiability, the analysis demonstrates that multiattribute instruments exhibit systematic and categorical non-possession of the principles required for quantitative meaning. Normalized logit values repeatedly collapse to the floor across independent axioms, indicating that these principles do not operate as binding constraints within the HTA knowledge base. The failures are structural, not technical. Multiattribute instruments deny unidimensionality by construction, lack true zero properties, permit negative values while claiming ratio status, substitute preference aggregation for measurement, and categorically exclude Rasch transformation for latent traits. As a result, arithmetic operations performed on their outputs are unlawful, and claims derived from them are insulated from empirical refutation.

The paper further shows that repair is impossible. Improvements in valuation methods, dimensional richness, statistical modeling, or Bayesian updating cannot resolve violations that are architectural in nature. Multiattribute instruments did not arise to solve a measurement problem but an administrative one: the need for a single summary number to support resource allocation. That substitution of preference for measurement has become entrenched across HTA institutions.

The paper concludes that multiattribute instruments must be abandoned as quantitative foundations for HTA. If this reconstruction does not occur, health technology assessment ceases to be a subject of interest to those concerned with the comparative impact of competing therapies. Only within such a framework can HTA recover scientific accountability, meet its duty of care to patients and clinicians, and re-enter the tradition of normal science in which claims are exposed to the risk of being wrong.

INTRODUCTION

For more than four decades, multiattribute utility instruments have occupied a central position in health technology assessment (HTA). Instruments such as the Health Utilities Index, EQ-5D, AQoL, and their variants have been treated as indispensable components of cost-utility analysis, supplying the utility weights required to construct quality-adjusted life years and, by extension, claims for cost-effectiveness. Their continued use is typically defended on pragmatic grounds: decision makers require a single summary metric, health is multidimensional, and no perfect

measure exists. These defenses have allowed multiattribute instruments to persist largely unquestioned as quantitative foundations for HTA.

This paper argues that this persistence can no longer be justified. The claim advanced here is not that multiattribute instruments are imperfect, controversial, or in need of refinement. It is that they are epistemically incompatible with measurement as understood in normal science. They do not, and cannot, generate quantities that satisfy the axioms of representational measurement. As a result, they cannot support lawful arithmetic, empirically falsifiable claims, or the cumulative evolution of objective knowledge. In this sense—and in this sense only—the appropriate scientific conclusion is that multiattribute instruments are dead as measurement instruments for HTA.

The "death" framing is deliberate and necessary. In science, death is declared not when a concept is unpopular, but when it is shown to be irreparable. Multiattribute instruments fail this test in an overdetermined way. They violate unidimensionality by construction, collapse qualitatively distinct attributes into single indices through preference weighting, lack true zero properties, permit negative values while claiming ratio status, and categorically exclude the only lawful transformation model for latent traits. These failures are independent, cumulative, and structural. No modification of scoring functions, valuation protocols, or model sophistication can repair them without dismantling the instruments themselves.

The persistence of multiattribute instruments has been sustained by a substitution that has gone largely unexamined: preference has been allowed to stand in for measurement. Utility theory has displaced representational measurement theory, enabling arithmetic operations on scores that do not represent empirical attributes. This substitution has been normalized through repetition, institutional endorsement, and journal publication, creating the appearance of quantitative rigor without the discipline that measurement requires. The result is a closed evaluative system in which numbers are authoritative but immune to refutation.

Recent work applying a fixed 24-item canonical diagnostic across HTA journals, national guidelines, and utility instruments makes this condition empirically visible. When interrogated against the axioms of measurement, falsifiability, and admissible arithmetic, multiattribute instruments consistently exhibit effective non-possession of the principles required for quantitative meaning. Normalized logit values collapse to the floor not because of analytical bias, but because these principles do not operate as binding constraints within the HTA knowledge base. The pattern is systematic across jurisdictions and institutions.

The implications are not merely theoretical. Claims derived from multiattribute instruments inform pricing, reimbursement, and access decisions that affect real patients and clinical practice. When such claims cannot be empirically tested or revised as quantitative claims, HTA fails its duty of care. Science advances by exposing claims to the risk of being wrong. Instruments that preclude that risk do not support learning; they enforce closure.

This paper therefore treats multiattribute instruments as a failed scientific object. Their death" is not rhetorical, but diagnostic. A framework that cannot generate falsifiable claims, cannot be empirically updated, and cannot distinguish error from uncertainty cannot serve as the quantitative foundation of HTA. What follows examines why multiattribute instruments were created, what

scientific modeling requires, what these instruments actually do, why repair is impossible, and what must abandon them if HTA is to recover its status as a field concerned with the real-world impact of competing therapies.

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

1. THE CREATION OF MULIATTRIBUTE INSTRUMENTS

Multiattribute instruments arose in health technology assessment not in response to a measurement problem posed by clinical science, but in response to an administrative problem posed by resource allocation. From the outset, the central challenge confronting early HTA was not how to measure health outcomes rigorously, but how to justify choices under budget constraint when competing technologies affected patients in different ways. Decision makers required a single summary quantity that would allow heterogeneous clinical effects to be compared, aggregated, and ranked. Multiattribute instruments emerged to supply that quantity.

The key historical move was the operationalization of "quality of life" as a composite object that could be assumed to be collapsed to a single number. Clinical science however does not present health as a single attribute; it presents outcomes as distinct phenomena, survival, symptom control, functional capacity, adverse events, and patient experience, each measured, if at all, on its own scale. None of these outcomes could, by itself, serve the purposes of cost-utility analysis, which required a generic outcome capable of spanning diseases, interventions, and populations. Rather than confronting this incompatibility directly, HTA reframed the problem. Health was reconceptualized as a bundle of attributes, and "quality of life" became the label under which those attributes could be combined.

This reconceptualization did not originate in measurement theory. It originated in welfare economics and decision analysis, where preference aggregation had long been used to collapse multiple attributes into a single utility index. That logic was imported wholesale into HTA. The question was no longer whether health outcomes could be measured as quantities, but whether preferences over health states could be elicited, weighted, and combined. Once that substitution was accepted, the need for measurement axioms receded. Utility scores did not need to represent empirical attributes; they needed only to support ranking, trade-offs, and aggregation for decision making.

Multiattribute instruments such as HUI, EQ-5D, and later AQoL were therefore designed as scoring systems rather than measurement systems. They classified health into domains, assigned levels within each domain, and used valuation surveys to attach preference weights to the resulting health states. The output was a single index, anchored by convention at full health and death, and readily usable in arithmetic. The success of these instruments lay not in their measurement

properties, but in their administrative usefulness. They delivered what HTA required: a number that could be multiplied by time, aggregated across individuals, and compared across interventions.

Crucially, this development occurred before representational measurement theory had any serious influence on HTA practice. The axioms governing unidimensionality, scale type, admissible transformations, and invariance were well established in the measurement literature, but they were not part of the intellectual toolkit of early HTA. The focus was on decision support, not on the conditions under which numbers legitimately represent attributes. As a result, the question "Is this a measure?" was largely displaced by the question "Is this useful for decisions?"

This displacement explains why multiattribute instruments were never designed to meet measurement axioms. They did not fail measurement accidentally; measurement was not their objective. Their objective was to enable cost-utility analysis under conditions of complexity and scarcity. The fact that health was multidimensional was treated as a reason to aggregate, not as a constraint on aggregation. The absence of a single clinical outcome was treated as a motivation to invent one, not as a warning sign about the limits of quantification.

Once embedded in HTA institutions, multiattribute instruments acquired a self-reinforcing legitimacy. Their outputs were incorporated into guidelines, textbooks, journals, and reimbursement processes, creating the appearance of a quantitative consensus. Over time, methodological refinement focused on improving valuation techniques, expanding descriptive domains, and enhancing sensitivity, not on revisiting the foundational question of measurement. The administrative problem had been solved, and the solution became taken for granted.

Understanding this origin is essential, because it clarifies the core thesis of this paper. Multiattribute instruments did not arise as failed attempts at measurement that might now be repaired. They arose as deliberate substitutes for measurement, optimized for decisional closure for new therapy options rather than empirical representation of their clinical impact. The subsequent application of representational measurement axioms does not reveal an unfortunate mismatch between theory and practice; it reveals that the instruments were never intended to satisfy those axioms. The "death" of multiattribute instruments in HTA therefore reflects not a change in standards, but the application of standards that were required for quantitative claims to be meaningful that had been recognized since Stevens' seminal and widely accepted 1946 classification of the scales of measurement where it was made clear only unidimensional interval and ratio measures supported arithmetic [1].

2. WHAT MEASUREMENT REQUIRES

Measurement in normal science is not a matter of convention, usefulness, or consensus. It is governed by axioms that specify when numbers legitimately represent empirical attributes and when arithmetic operations on those numbers are meaningful. These axioms are not discipline-specific preferences; they are the logical conditions that make quantification possible. Any framework that claims to produce quantitative evidence must satisfy them, irrespective of context or policy need; formalized in 1971 they were rapidly accepted as the only standards for meaningful measurement [2].

At the foundation of representational measurement is unidimensionality. A measure must represent variation along a single attribute. If multiple attributes are involved, they must be measured separately. Collapsing qualitatively distinct attributes into a single index does not create a measure; it creates a composite score. Without unidimensionality, there is no coherent empirical structure for numbers to represent, and arithmetic operations lack meaning.

A second requirement is invariance. The meaning of a measure must be stable across persons, contexts, and comparisons. Differences must represent the same empirical difference wherever they occur. Without invariance, comparisons cannot be interpreted, and aggregation is indefensible. Invariance is not an optional refinement; it is the condition that allows measurement to support generalization and learning.

Measurement further requires adherence to admissible transformations, which define what numerical operations preserve empirical meaning. These transformations determine scale type. Nominal scales permit classification only. Ordinal scales permit ordering but not arithmetic on differences. Interval scales permit addition and subtraction because equal differences correspond to equal empirical differences, but they lack a true zero. Ratio scales alone possess a true zero representing the absence of the attribute and therefore permit multiplication, division, and meaningful ratios. Arithmetic operations are not universally permissible; they are conditional on scale type. Treating ordinal or interval quantities as ratio measures is not a mild approximation; it is a violation of measurement law.

The requirement of a true zero is especially critical. A true zero is not a naming convention or an anchor point; it is an empirical property indicating the absence of the attribute being measured. Without a true zero, ratios are meaningless. Negative values on a purported ratio scale, such as negative utilities for states worse than death, are a logical impossibility. Any framework that multiplies quantities lacking a true zero commits an irreparable arithmetic error.

These requirements apply equally to latent traits, with an additional constraint. Latent traits, such as need fulfillment, functioning, or symptom burden, are not directly observable. Their measurement requires a transformation from ordinal observations to a scale with interval or ratio properties. In normal science, this transformation is governed by explicit models that impose unidimensionality and test invariance empirically.

The Rasch model occupies a unique position here [3 4]. It is not a psychometric option among many; it is the only transformation model that satisfies the axioms of representational measurement for latent traits. Rasch measurement simultaneously estimates item difficulty and person ability on a common logit scale, enforcing unidimensionality and invariance as testable conditions rather than assumptions. When data fit the Rasch model, the resulting logit scale supports meaningful comparisons of differences and ratios. When data do not fit, measurement fails. There is no fallback position in which ordinal scores can be treated as quantities by convention. The key Rasch output is possession of a latent trait on a logit ratio scale. This is the only framework for assessing therapy impact on latent traits

This point cannot be softened. Without Rasch or an equivalent conjoint measurement model, ordinal or observational responses remain ordinal. Summation, averaging, weighting, or scaling of

ordinal responses does not create measurement. It produces scores that are numerically manipulable but empirically meaningless. Summed responses to Likert question are ordinal; they cannot support arithmetic operations.

These axioms define the benchmark against which all HTA instruments and claims must be judged. They do not bend to administrative convenience, policy urgency, or institutional precedent. A claim that violates them is not "less precise" or "imperfect"; it is just not a quantitative claim at all. Measurement either exists or it does not. The purpose of the sections that follow is not to renegotiate these requirements, but to apply them.

3. THE ROLE OF MULTIATTRIBUTE INSTRUMENTS

Multiattribute instruments such as the Health Utilities Index, EQ-5D, AQoL, and related systems share a common internal logic, regardless of differences in descriptive detail or valuation technique. They do not attempt to measure a single empirical attribute. Instead, they decompose health into multiple, qualitatively distinct attributes and then reassemble those attributes into a single numerical index using preference-based scoring rules. This architecture defines what these instruments do and, equally important, what they cannot do. This architecture denies measurement.

The process begins with classification rather than measurement. Health is partitioned into domains to support health state descriptions: mobility, pain, mental health, self-care, cognition, social functioning, and similar attributes. Each domain is represented by ordered response categories. At this stage, the instrument produces a multidimensional descriptive profile. Nothing resembling measurement has yet occurred. The responses are ordinal, domain-specific, and incommensurable across attributes. There is no concept of scales of measurement.

The critical move occurs at the next step: population samples are asked to express preferences over hypothetical health states defined by combinations of attribute levels. These preferences are then used to construct a scoring algorithm, additive, multiplicative, or hybrid, that assigns a single number to each health state. This number is typically anchored at "full health" and "dead," with negative values permitted for states judged worse than death. The resulting index is labeled a "utility" and treated as if it were a quantity.

This process substitutes preference aggregation for measurement. Preferences are relational judgments about desirability; they do not represent empirical attributes. Aggregating preferences does not create a measure any more than voting creates temperature. The scoring function may be appear mathematically sophisticated, but sophistication does not alter category. No empirical structure is specified to which the numbers are homomorphic. The index represents a negotiated valuation, not an attribute possessed by individuals.

Proponents often argue that anchoring the scale at dead = 0 and full health = 1 creates a meaningful zero and justifies arithmetic. This argument fails on first principles. A true zero is not an anchor point; it is an empirical condition representing the absence of the attribute being measured. Death is not the absence of "quality of life" in any measurable sense, nor is it commensurate with gradations of living health states. Allowing negative values further confirms the absence of ratio

properties. A scale that permits negative quantities cannot be a ratio scale. Anchoring and normalization are conventions; they do not confer measurement status.

Similarly, replacing simple summation with complex utility functions does not rescue measurement. Whether attributes are combined additively or multiplicatively is irrelevant to the core issue. Without unidimensionality and invariance, no functional form can create a measure. Complexity here serves a rhetorical function: it obscures the absence of measurement behind mathematical formalism. The output may be numerically precise, but precision is not meaning.

Crucially, multiattribute instruments do not test whether their scores satisfy the conditions required for measurement. They do not impose unidimensionality as a requirement, because their design explicitly denies it. They do not test invariance, because there is no single attribute whose meaning could remain invariant. They do not specify admissible transformations, because the scale type is assumed rather than demonstrated. Arithmetic is performed because it is needed, not because it is licensed.

The categorical exclusion of Rasch measurement is decisive here. Rasch would require the instrument to define a single latent trait and to demonstrate that items function invariantly along that trait. Multiattribute instruments cannot meet this requirement without abandoning their defining architecture. As a result, Rasch is not merely unused; it is irrelevant by design and forgotten. Measurement is redefined as preference scoring, and the axioms that would otherwise govern transformation are displaced.

The consequence is that multiattribute instruments are scoring systems, not measurement systems. They generate indices suitable for ranking, aggregation, and modeling, but those indices do not represent quantities. They cannot support meaningful multiplication by time, aggregation across persons, or ratio comparisons, because the numbers lack the properties required for such operations. Calling these outputs “utilities” or “measures” does not change their status.

This distinction matters because the language of measurement and the formal axioms carry scientific authority across multiple disciplines. By presenting preference-based scores as quantities, multiattribute instruments borrow the legitimacy of measurement without submitting to its discipline. The result is an evaluative framework in which impossible arithmetic proceeds unchallenged, even though the numbers to which it is applied do not measure anything. This is not a minor methodological weakness. It is the defining characteristic of the multiattribute approach and the reason it cannot serve as a quantitative foundation for health technology assessment. It’s not science. It’s not measurement. It’s 40 years of numerical storytelling.

4. WHAT THE LOGIT EVIDENCE SHOWS ACROSS COUNTRIES AND JOURNALS

The application of the 24-item canonical diagnostic across HTA journals, national guidelines, and multiattribute instruments yields a strikingly consistent empirical pattern. Regardless of jurisdiction, institutional level, or publication venue, the same foundational propositions repeatedly collapse to the floor of the normalized logit scale. These results are not sporadic, marginal, or sensitive to analytical choices. They represent systematic non-possession of the

principles required for measurement, lawful arithmetic, and falsifiable quantitative claims within the HTA knowledge base.

Across international journals such as the *International Journal of Technology Assessment in Health Care* and *Value in Health,* national publication vehicles such as the *Canadian Journal of Health Technologies*, the logit profiles exhibit a shared structure. Statements asserting that measures must be unidimensional, that multiplication requires a ratio measure, that measurement must precede arithmetic, and that Rasch rules are required to transform ordinal responses into interval scales routinely register at logit values of −2.50 or adjacent floor levels, indicating effective non-possession within the evaluated knowledge base. These propositions are not controversial within measurement science; they are axiomatic. Their categorical exclusion indicates that they do not operate as binding constraints on what counts as an admissible quantitative claim in HTA.

The same pattern appears in assessments of national HTA guidelines. Whether examining European, North American, or Australasian frameworks, the results converge. While there may be variation in rhetoric, process detail, or modeling sophistication, the canonical diagnostics reveal that representational measurement axioms are not enforced anywhere as conditions of admissibility. Arithmetic proceeds without prior demonstration of scale type. Composite indices are accepted without unidimensionality. Simulation outputs are treated as evidence despite lacking falsifiability. The knowledge base is consistent in what it excludes.

Most revealing is the alignment between institutional outputs and the instruments on which they rely. When the canonical diagnostic is applied to multiattribute instruments such as HUI and AQoL, the same propositions collapse to the floor. Unidimensionality is categorically absent. Rasch transformation is excluded. Ratio properties are denied while ratio arithmetic is performed. Falsifiability is displaced by model recalibration. The coincidence is not accidental. Journals and guidelines do not merely tolerate these failures; they institutionalize them by adopting instruments whose architecture makes compliance with measurement axioms impossible.

The repeated appearance of −2.50 values warrants more explicit interpretation. In this framework, −2.50 does not signify ignorance or disagreement. It denotes effective non-possession: the proposition does not function as a binding principle within the evaluated knowledge base, irrespective of whether it is occasionally acknowledged in narrative form. A proposition at −2.50 cannot constrain practice. It cannot disqualify claims. It cannot generate refutation. Its presence in discourse, if any, is inert.

Importantly, the failures observed are independent and cumulative. Multiattribute instruments do not merely violate one axiom; they violate several, each sufficient on its own to invalidate quantitative claims. Textbooks, journals and guidelines, in turn, reflect and reinforce these violations by accepting arithmetic operations that depend on those axioms while excluding the axioms themselves. This overdetermination matters. It rules out the possibility that the problem is one of local misinterpretation, poor training, or inconsistent application. The pattern is stable across countries, agencies, and publication venues. It is witnessed globally.

The logit evidence therefore establishes that the continued use of multiattribute instruments is not a matter of unresolved debate. It is a matter of entrenched global epistemic practice. The HTA knowledge base has evolved in a way that systematically excludes the conditions required for measurement while retaining the appearance of quantification. This exclusion is not remedied by methodological refinement, because refinement occurs within the same closed framework.

Seen in this light, the "death" of multiattribute instruments is not a rhetorical flourish. It is an empirical diagnosis. Instruments that cannot, even in principle, satisfy the axioms of measurement are not candidates for incremental repair. When the same diagnostic applied across instruments, journals, and guidelines yields the same pattern of categorical non-possession, the conclusion follows that the failure is structural. The logit evidence shows that multiattribute instruments are not temporarily deficient components of HTA; they are incompatible with the epistemic requirements of quantitative science.

5. WHY REPAIR IS IMPOSSIBLE

The continued defense of multiattribute instruments in health technology assessment rests almost entirely on the belief that their defects are technical rather than architectural. It is routinely asserted that problems of validity, sensitivity, or interpretability can be addressed through better preference weights, more descriptive dimensions, improved valuation surveys, Bayesian updating, or more sophisticated modeling. This belief is mistaken. The failures of multiattribute instruments are not contingent shortcomings awaiting refinement; they are necessary consequences of the design logic itself. Repair is impossible because there is nothing to repair without abandoning the defining features of the instruments.

The core architectural flaw is the rejection of unidimensionality. Multiattribute instruments begin by decomposing health into multiple, qualitatively distinct attributes and then collapsing those attributes into a single index. This is not an incidental modeling choice; it is the essence of the approach. No increase in the number of dimensions can correct this defect. Adding dimensions increases descriptive richness while further entrenching dimensional heterogeneity. A composite constructed from more attributes is not closer to unidimensional measurement; it is further from it. Because unidimensionality is a necessary condition for measurement, any instrument that denies it by construction is irreparable as a measurement system. A requirement recognized for centuries in the physical sciences.

A second irreparable defect is the absence of a true zero. Multiattribute utilities are anchored by convention, typically at "full health" and "dead," with negative values permitted for states judged worse than death. No adjustment to anchoring, scaling, or normalization can transform this convention into a true zero. A true zero is an empirical property representing the absence of the attribute, not a boundary chosen for convenience. Allowing negative values is not a minor anomaly; it is a logical disqualification of ratio status. Because multiplication requires a ratio scale, any framework that depends on multiplying such utilities by time cannot be repaired without abandoning the multiplication itself.

Attempts to rescue multiattribute instruments through improved weighting schemes misunderstand the nature of the problem. Preference weights, whether elicited through time trade-off, standard

gamble, discrete choice experiments, or hybrid methods, are expressions of relative desirability. They do not represent empirical magnitudes. Refining elicitation protocols may improve consistency or face validity, but it cannot create measurement. No amount of statistical sophistication can convert preferences into quantities without specifying and testing an empirical structure to which the numbers correspond.

Similarly, replacing simple summation with complex utility functions does not address the underlying violation. Whether attributes are combined additively, multiplicatively, or through interaction terms is irrelevant in the absence of unidimensionality and invariance. Complex mathematics can obscure the problem, but it cannot resolve it. The repeated historical pattern is instructive: each methodological "advance" in multiattribute instruments has increased mathematical complexity while leaving the foundational measurement violations untouched. This is not progress; it is elaboration within a closed framework.

Bayesian updating and probabilistic modeling fare no better. Updating beliefs about parameter values does not create a measure where none exists. Bayesian methods operate on probabilities; they do not confer scale properties on the quantities being modeled. If the underlying numbers lack admissible transformations, updating them merely propagates error more efficiently. Likewise, sensitivity analysis cannot rescue non-measurement. Demonstrating robustness to assumptions does not establish empirical meaning. A claim can be robustly wrong.

The categorical exclusion of Rasch measurement is decisive evidence that repair is impossible. Rasch measurement would require multiattribute instruments to specify a single latent trait and to demonstrate invariant item functioning along that trait. Compliance with Rasch would therefore require abandoning the multiattribute architecture altogether. The fact that Rasch is consistently rejected, not debated, but excluded shows that the HTA knowledge base understands, at least implicitly, that adopting measurement discipline would dismantle the utility framework. Repair is impossible because repair would require transformation into something else.

These failures are overdetermined. Multiattribute instruments violate unidimensionality, lack a true zero, deny ratio properties while performing ratio arithmetic, exclude lawful transformation models, and insulate claims from falsification through simulation. Each violation is independently sufficient to invalidate quantitative claims. Together, they make the framework beyond repair. No single adjustment can resolve all failures simultaneously, and resolving any one would require abandoning the core purpose of the instrument.

This overdetermination is why reformist defenses persist rhetorically but fail analytically. It is always possible to argue that no measure is perfect, that decisions must still be made, or that some information is better than none. These arguments are administrative, not scientific. They concede the impossibility of measurement while insisting on arithmetic anyway. Once that concession is made, the death of multiattribute instruments as measurement tools is already acknowledged in substance, if not in name.

6. WHAT REPLACES MULIATTRIBUTE INSTRUMENTS

If multiattribute instruments are abandoned, health technology assessment does not lose quantitative rigor; it gains it. The replacement is not another composite index or a more elaborate preference system, but a fundamentally different evaluative architecture: a portfolio of single-attribute, empirically evaluable claims, each supported by measurement-valid scales and explicit protocols. This shift replaces decisional closure with scientific accountability.

At the core of this architecture is the single-claim principle. Each value claim must correspond to a single attribute of interest, clearly defined and measured independently. For manifest outcomes such as survival, time to event, hospital days avoided, or resource use this requires linear ratio measures with a true zero and invariant meaning. Ratio scales permit lawful arithmetic, meaningful ratios, and transparent aggregation where appropriate. Claims based on such measures are not model artifacts; they are empirically testable statements about observable events.

Latent constructs require a different but equally rigorous approach. Attributes such as symptom burden, functional capacity, or need fulfillment cannot be observed directly and therefore cannot be measured by counting or timing. For these constructs, the only admissible quantitative framework is Rasch measurement, which yields a logit ratio scale under explicit axioms of unidimensionality and invariance. Rasch measurement is not a psychometric convenience; it is the sole transformation model that converts ordinal observations into a scale capable of supporting meaningful comparison. When data do not fit the Rasch model, measurement fails, and the claim must be revised or abandoned.

Together, linear ratio measures for manifest outcomes and Rasch logit ratio measures for latent traits define the only quantitative space in which HTA claims can be meaningful. No arithmetic is permitted outside this space. Summation, multiplication, or aggregation across attributes is replaced by parallel evaluation: multiple claims assessed independently rather than collapsed into a single index. This preserves scientific integrity while allowing decision makers to consider trade-offs explicitly rather than hiding them inside composite scores.

This framework naturally leads to a portfolio of claims rather than a single synthetic outcome. A product submission may include separate claims for survival, hospitalization, symptom improvement, functional change, adherence, or other relevant outcomes, each with its own measurement standard, target population, and evaluation protocol. The portfolio does not require dimensional homogeneity because it does not force incommensurable attributes into a single number. Comparison occurs at the level of evidence, not arithmetic.

Crucially, each claim in the portfolio is empirically vulnerable. Protocols specify how claims will be evaluated, when results will be reported, and under what conditions claims will be rejected or revised. This restores falsification as an operational principle. Claims are not insulated by lifetime simulations or reference cases; they are exposed to real-world performance. Learning becomes possible.

This replacement architecture also restores duty of care. Clinicians receive claims grounded in observable outcomes or measurement-valid latent traits, not preference-weighted composites. Patients are represented by attributes they actually possess, not by abstract social valuations of

hypothetical health states. Health systems gain information that can be updated as evidence accumulates rather than locked into permanent closure.

Abandoning multiattribute instruments does not simplify HTA; it disciplines it. It replaces the illusion of precision with lawful measurement, and decisional convenience with scientific accountability. A portfolio of single-attribute claims, supported by linear ratio and Rasch logit ratio measures, is not an optional reform. It is the only quantitative framework consistent with the axioms of measurement and the requirements of normal science.

CONCLUSION

This paper has argued not that multiattribute utility instruments have reached the end of their scientific life in health technology assessment, but that they never possessed one. From their inception, they failed to meet the axioms of representational measurement required for quantitative claims. Their continued use reflects institutional entrenchment rather than scientific legitimacy; they were dead on arrival.

The companion Logit review of the reference case simulation makes clear it did not arise to correct these failures, but to preserve them. It provides the analytical environment in which non-measures can be combined, projected, and insulated from empirical refutation. In doing so, it converts foundational measurement errors into stable policy artifacts. The survival of multiattribute utilities and the survival of the reference case are therefore inseparable: both persist not because they work scientifically, but because together they enforce closure.

This conclusion does not rest on preference, policy disagreement, or methodological fashion. It follows directly from the axioms of representational measurement and is confirmed empirically by canonical logit interrogation across instruments, journals, and national HTA frameworks. Multiattribute instruments do not, and cannot, generate quantities that support lawful arithmetic, empirical falsification, or the evolution of objective knowledge.

The failure is not singular but overdetermined. Multiattribute instruments deny unidimensionality by construction, lack true zero properties, permit negative values while claiming ratio status, exclude Rasch transformation for latent traits, and insulate claims from empirical refutation through simulation. Each failure is independently sufficient to invalidate quantitative claims; together they render repair conceptually impossible. Refinement, complexity, and improved valuation cannot resolve a category error.

The persistence of multiattribute instruments reflects institutional convenience rather than scientific legitimacy. They arose to solve an administrative problem, producing a single number for resource allocation, not a measurement problem. Over time, preference aggregation was substituted for measurement, and decisional closure was mistaken for evidence. The canonical logit evidence shows that this substitution has become entrenched across HTA, not contested.

Declaring the death of multiattribute instruments is therefore not provocative; it is corrective. It marks the point at which HTA must choose between continuing to generate authoritative but non-evaluable numbers, or reconstructing itself around measurement-valid, falsifiable claims. That

reconstruction requires abandoning composite indices and adopting a portfolio of single-attribute claims supported by linear ratio measures for manifest outcomes and Rasch logit ratio measures for latent traits.

If this paradigm shift does not occur, health technology assessment ceases to be a subject of interest to those concerned with the comparative impact of competing therapies. Only within such a framework can HTA recover scientific accountability, meet its duty of care to patients and clinicians, and re-enter the tradition of normal science in which claims are exposed to the risk of being wrong.

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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