RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

THE MEASUREMENT FUTURE IN HEALTH TECHNOLGY ASSESSMENT

ITALY: MEASUREMENT FAILURE IN LEGACY HEALTH TECHNOLOGY ASSESSMENT

Paul C Langley PhD Adjunct Professor, College of Pharmacy, University of Minnesota, Minneapolis, MN

LOGIT WORKING PAPER No 1075 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

Background: *Contemporary health technology assessment (HTA) depends upon extensive arithmetic involving preference-based utilities, quality-adjusted life years (QALYs), cost-effectiveness ratios and decision models. Representational measurement theory requires a logically prior question: have the quantities entering these operations been established as measures possessing the scale properties required for the arithmetic performed upon them? This paper examines whether the Italian national HTA knowledge base recognizes this requirement or exhibits measurement inversion, where arithmetic is accepted before its measurement foundations have been established.*

Methods: *A large language model was used to interrogate the publicly identifiable Italian HTA knowledge environment, including national institutional guidance, methodological conventions, analytical practices and the wider European HTA framework within which Italy operates. The interrogation employed 24 canonical TRUE/FALSE propositions derived from representational measurement theory and Rasch measurement principles. Categorical endorsement probabilities were assigned to characterize the extent to which the knowledge base explicitly or implicitly reinforces each proposition and were transformed to normalized logits.*

Results: *The results demonstrate a systematic inversion between measurement requirements and established HTA practice. The propositions that measurement precedes arithmetic and that representational measurement axioms are required for arithmetic each received endorsement probabilities of 0.10 (normalized logit −2.20), while multiplication requiring ratio measurement received 0.15 (−1.95). In contrast, false propositions required to sustain legacy HTA were strongly reinforced: the QALY as a ratio measure received 0.90 (+2.20), QALY aggregation 0.90 (+2.20), and reference-case simulations as generating falsifiable claims 0.90 (+2.20). Principles required for Rasch measurement of latent attributes were essentially absent, several receiving probabilities of 0.05 (−2.50).*

Conclusions: *The Italian profile is consistent with structural measurement inversion rather than an isolated methodological omission. Italy operates within a sophisticated but inherited international HTA framework in which quantitative convention has become institutionalized without demonstrated measurement foundations. European harmonization and widespread methodological acceptance cannot establish scientific admissibility. The challenge is therefore one of institutional responsibility. Italian agencies, universities and research groups need not wait for international consensus to begin replacing measurement inversion with an analytical framework in which the sequence is restored: attribute, measurement, admissible arithmetic, empirically evaluable claims, replication and falsification.*

INTRODUCTION

Health technology assessment in Italy operates within a mature and increasingly coordinated national and European environment. Pharmaceutical pricing and reimbursement, medical-device assessment, regional decision-making, university research and health-economic evaluation draw upon an extensive body of methods developed over several decades. Cost-effectiveness analysis, preference-based utilities, quality-adjusted life years (QALYs), decision modelling, evidence synthesis and uncertainty analysis have become familiar components of this environment. Their continued use reflects Italy's participation in the wider European development of contemporary or legacy HTA and its commitment to increasingly systematic approaches to healthcare decision-making.

The scientific question addressed in this paper, however, precedes all of these methods: were the quantities entering the arithmetic established as measures before the arithmetic was undertaken?

This question defines the problem of measurement inversion. In normal quantitative science, measurement precedes arithmetic. The attribute must first be identified, the properties of the proposed measure established, the measurement scale demonstrated and the transformations and arithmetic admissible for that scale determined [1]. Only then can numerical operations acquire empirical meaning. Measurement inversion reverses this sequence. The analytical operation is accepted first and the numerical inputs are subsequently treated as though they possess whatever measurement properties the operation requires. Arithmetic no longer follows from measurement; presumed measurement properties follow from the arithmetic that analysts wish to perform.

The distinction is fundamental for HTA. The QALY requires multiplication of a health-state value by time. Cost-per-QALY analysis subsequently requires subtraction and division. Decision models may then aggregate, average, discount and probabilistically manipulate these quantities over extended simulated time horizons. None of these operations establishes the measurement properties of its inputs. If the required properties were absent before the arithmetic began, increasing the sophistication of the subsequent analysis cannot supply them. A measurement failure cannot be repaired by additional arithmetic [2].

Italy provides an important setting in which to examine this inversion. Its national HTA coordination knowledge base incorporates the activities and methodological influence of institutions including AIFA, AGENAS and the Istituto Superiore di Sanità, while simultaneously operating within the broader methodological environment of European HTA. The resulting system is neither analytically primitive nor methodologically undeveloped. On the contrary, it represents a substantial infrastructure for evidence evaluation, economic analysis and healthcare decision-making. The issue is whether representational measurement provides the scientific foundation governing that infrastructure.

This distinction matters because Italy did not develop contemporary HTA in intellectual isolation. Its methods reflect the international diffusion of an established analytical framework in which utilities, QALYs, cost-effectiveness analysis and reference-case modelling had already acquired methodological legitimacy. Alignment with European and international practice can therefore provide procedural consistency without answering the logically prior measurement question. A method does not become scientifically admissible because it is widely used, incorporated into guidance or accepted across jurisdictions. International consensus can transmit a measurement failure just as effectively as it can transmit a scientific advance.

The Italian national knowledge base is therefore interrogated here not to determine whether particular Italian researchers understand measurement theory, nor to attribute beliefs to individual practitioners. The unit of analysis is the knowledge environment itself: the published guidance, methodological conventions, analytical practices and conceptual assumptions through which HTA is represented and reproduced. The large language model interrogation estimates the extent to which this knowledge base endorses, implies or reinforces 24 canonical propositions concerning representational measurement, scale properties, admissible arithmetic, QALYs, Rasch measurement, latent attributes and falsifiability. The objective is to identify what the knowledge system appears to possess, what it appears not to possess and, critically, whether its implicit commitments are consistent with the requirements of lawful quantitative measurement.

The Italian results raise a more serious possibility than the omission of measurement theory from an otherwise defensible analytical framework. They suggest that the omission may be structural. Foundational propositions that should constrain quantitative operations receive weak endorsement, while propositions incompatible with representational measurement but necessary to sustain familiar legacy HTA practices receive strong reinforcement. Measurement inversion, on this interpretation, is not an isolated technical error. It has become embedded within the epistemic architecture of HTA.

Italy therefore provides a national test of a much larger international question. Has legacy HTA created a quantitative system in which arithmetic became institutionalized before its measurement foundations were established? If so, the challenge facing Italian HTA is not simply to improve existing models or refine economic-evaluation methods. It is to restore the scientific sequence that should have governed quantitative therapy evaluation from the outset: attribute first, measurement second, admissible arithmetic third, and empirical claims only thereafter.

INTERROGATING THE LARGE LANGUAGE MODEL

A large language model **(**LLM**)** is an artificial intelligence system designed to understand, generate, and manipulate human language by learning patterns from vast amounts of text data. Built on deep neural network architectures, most commonly transformers, LLMs analyze relationships between words, sentences, and concepts to produce contextually relevant responses. During training, the model processes billions of examples, enabling it to learn grammar, facts, reasoning patterns, and even subtle linguistic nuances. Once trained, an LLM can perform a wide range of tasks: answering questions, summarizing documents, generating creative writing, translating languages, assisting with coding, and more. Although LLMs do not possess consciousness or true understanding, they simulate comprehension by predicting the most likely continuation of text based on learned patterns. Their capabilities make them powerful tools for communication, research, automation, and decision support, but they also require careful oversight to ensure accuracy, fairness, privacy, and responsible use

In this Logit Working Paper, "interrogation" refers not to discovering what an LLM *believes,* it has no beliefs, but to probing the content of the *corpus-defined knowledge space* we choose to analyze. This knowledge base is enhanced if it is backed by accumulated memory from the user. In this case the interrogation relies also on 12 months of HTA memory from continued application of the system to evaluate HTA experience. The corpus is defined before interrogation: it may consist of a journal (e.g., *Value in Health*), a national HTA body, a specific methodological framework, or a collection of policy documents. Once the boundaries of that corpus are established, the LLM is used to estimate the conceptual footprint within it. This approach allows us to determine which principles are articulated, neglected, misunderstood, or systematically reinforced.

In this HTA assessment, the objective is precise: to determine the extent to which a given HTA knowledge base or corpus, global, national, institutional, or journal-specific, recognizes and reinforces the foundational principles of representational measurement theory (RMT). The core principle under investigation is that measurement precedes arithmetic; no construct may be treated as a number or subjected to mathematical operations unless the axioms of measurement are satisfied. These axioms include unidimensionality, scale-type distinctions, invariance, additivity, and the requirement that ordinal responses cannot lawfully be transformed into interval or ratio quantities except under Rasch measurement rules.

The HTA knowledge space is defined pragmatically and operationally. For each jurisdiction, organization, or journal, the corpus consists of:

- published HTA guidelines
- agency decision frameworks
- cost-effectiveness reference cases
- academic journals and textbooks associated with HTA
- modelling templates, technical reports, and task-force recommendations
- teaching materials, methodological articles, and institutional white papers

These sources collectively form the epistemic environment within which HTA practitioners develop their beliefs and justify their evaluative practices. The boundary of interrogation is thus not the whole of medicine, economics, or public policy, but the specific textual ecosystem that sustains HTA reasoning. . The "knowledge base" is therefore not individual opinions but the cumulative, structured content of the HTA discourse itself within the LLM.

THE ITALIAN NATIONAL HTA KNOWLEDGE BASE

The Italian national HTA knowledge base is defined here as the publicly identifiable body of guidance, methodological conventions, institutional practices and analytical assumptions that establishes the framework within which health technology assessment is undertaken, communicated and reproduced in Italy. It is not intended to represent the beliefs or knowledge of individual researchers, analysts or decision-makers. The object of interrogation is the collective epistemic environment of Italian HTA.

At the national level, this environment is shaped principally by the Agenzia Italiana del Farmaco (AIFA), the Agenzia Nazionale per i Servizi Sanitari Regionali (AGENAS), the Ministry of Health and the Istituto Superiore di Sanità, together with the methodological literature, economic-evaluation practices and wider European HTA conventions upon which Italian assessment draws. These institutions perform different functions, but collectively they contribute to the standards through which evidence is generated, assessed and interpreted for pharmaceutical reimbursement, medical technologies and healthcare decision-making.

The knowledge base therefore extends beyond formal methodological guidance. It encompasses the recurring quantitative constructs and analytical practices through which Italian HTA operates, including economic evaluation, cost-effectiveness and cost-utility analysis, preference-based utilities, QALYs, evidence synthesis, decision modelling and uncertainty analysis. These practices provide evidence of the measurement assumptions that the system implicitly accepts even where those assumptions are not explicitly stated.

This distinction is essential to the interpretation of the interrogation. A knowledge base may behave as though a proposition is true without any institution formally asserting that proposition. For example, repeated multiplication of preference scores by time does not require an explicit declaration that the preference score possesses the measurement properties required for multiplication. The operation itself reveals that those properties are being presumed. Similarly, aggregation of QALYs, calculation of incremental QALYs and division of incremental costs by incremental QALYs demonstrate reliance upon particular measurement assumptions whether or not those assumptions have ever been formally examined.

The Italian knowledge base must also be understood within its European context. Italian HTA has developed through participation in a wider international methodological environment in which utilities, QALYs, cost-effectiveness analysis and reference-case modelling were already established components of accepted practice. Alignment with these conventions provides methodological consistency across jurisdictions, but it does not independently establish the measurement admissibility of the quantities and operations employed. The Italian knowledge base may therefore reproduce measurement assumptions inherited from the wider HTA environment rather than assumptions developed or explicitly defended within Italy itself.

The interrogation that follows should consequently be interpreted as an assessment of knowledge-system possession. It asks whether the Italian national HTA environment recognizably possesses and reinforces the principles required by representational measurement and, equally importantly, whether its established practices reinforce propositions incompatible with those principles. The resulting endorsement probabilities characterize the conceptual structure of the knowledge base. They are not survey estimates, measures of individual competence or claims about what particular Italian HTA practitioners personally believe.

This national knowledge base provides the appropriate level at which to investigate measurement inversion. If the principles that should determine admissible arithmetic are weakly represented while the assumptions required to sustain legacy HTA arithmetic are strongly reinforced, the problem is no longer adequately described as an isolated methodological omission. It indicates an epistemic structure in which measurement inversion has become embedded in the normal operation and reproduction of HTA.

CATEGORICAL PROBABILITIES

In the present application, the interrogation is tightly bounded. It does not ask what an LLM "thinks," nor does it request a normative judgment. Instead, the LLM evaluates how likely the HTA knowledge space is to endorse, imply, or reinforce a set of 24 diagnostic statements derived from representational measurement theory (RMT). Each statement is objectively TRUE or FALSE under RMT. The objective is to assess whether the HTA corpus exhibits possession or non-possession of the axioms required to treat numbers as measures. The interrogation creates a categorical endorsement probability: the estimated likelihood that the HTA knowledge base endorses the statement whether it is true or false; *explicitly or implicitly.*

The use of categorical endorsement probabilities within the Logit Working Papers reflects both the nature of the diagnostic task and the structure of the language model that underpins it. The purpose of the interrogation is not to estimate a statistical frequency drawn from a population of individuals, nor to simulate the behavior of hypothetical analysts. Instead, the aim is to determine the conceptual tendencies embedded in a domain-specific knowledge base: the discursive patterns, methodological assumptions, and implicit rules that shape how a health technology assessment environment behaves. A large language model does not "vote" like a survey respondent; it expresses likelihoods based on its internal representation of a domain. In this context, endorsement probabilities capture the strength with which the knowledge base, as represented within the model, supports a particular proposition. Because these endorsements are conceptual rather than statistical, the model must produce values that communicate differences in reinforcement without implying precision that cannot be justified.

This is why categorical probabilities are essential. Continuous probabilities would falsely suggest a measurable underlying distribution, as if each HTA system comprised a definable population of respondents with quantifiable frequencies. But large language models do not operate on that level. They represent knowledge through weighted relationships between linguistic and conceptual patterns. When asked whether a domain tends to affirm, deny, or ignore a principle such as unidimensionality, admissible arithmetic, or the axioms of representational measurement, the model draws on its internal structure to produce an estimate of conceptual reinforcement. The precision of that estimate must match the nature of the task. Categorical probabilities therefore provide a disciplined and interpretable way of capturing reinforcement strength while avoiding the illusion of statistical granularity.

The categories used, values such as 0.05, 0.10, 0.20, 0.50, 0.75, 0.80, and 0.85, are not arbitrary. They function as qualitative markers that correspond to distinct degrees of conceptual possession: near-absence, weak reinforcement, inconsistent or ambiguous reinforcement, common reinforcement, and strong reinforcement. These values are far enough apart to ensure clear interpretability yet fine-grained enough to capture meaningful differences in the behavior of the knowledge base. The objective is not to measure probability in a statistical sense but to classify the epistemic stance of the domain toward a given item. A probability of 0.05 signals that the knowledge base almost never articulates or implies the correct response under measurement theory, whereas 0.85 indicates that the domain routinely reinforces it. Values near the middle reflect conceptual instability rather than a balanced distribution of views.

Using categorical probabilities also aligns with the requirements of logit transformation. Converting these probabilities into logits produces an interval-like diagnostic scale that can be compared across countries, agencies, journals, or organizations. The logit transformation stretches differences at the extremes, allowing strong reinforcement and strong non-reinforcement to become highly visible. Normalizing logits to the fixed ±2.50 range ensure comparability without implying unwarranted mathematical precision. Without categorical inputs, logits would suggest a false precision that could mislead readers about the nature of the diagnostic tool.

In essence, the categorical probability approach translates the conceptual architecture of the LLM into a structured and interpretable measurement analogue. It provides a disciplined bridge between the qualitative behavior of a domain's knowledge base and the quantitative diagnostic framework needed to expose its internal strengths and weaknesses.

The LLM computes these categorical probabilities from three sources:

1. **Structural content of HTA discourse**
   If the literature repeatedly uses ordinal utilities as interval measures, multiplies non-quantities, aggregates QALYs, or treats simulations as falsifiable, the model infers high reinforcement of these false statements.
2. **Conceptual visibility of measurement axioms**
   If ideas such as unidimensionality, dimensional homogeneity, scale-type integrity, or Rasch transformation rarely appear, or are contradicted by practice, the model assigns low endorsement probabilities to TRUE statements.
3. **The model's learned representation of domain stability**
   Where discourse is fragmented, contradictory, or conceptually hollow, the model avoids assigning high probabilities. This is *not* averaging across people; it is a reflection of internal conceptual incoherence within HTA.

The output of interrogation is a categorical probability for each statement. Probabilities are then transformed into logits [$\ln(p/(1-p))$], capped to ±4.0 logits to avoid extreme distortions, and normalized to ±2.50 logits for comparability across countries. A positive normalized logit indicates reinforcement in the knowledge base. A negative logit indicates weak reinforcement or conceptual absence. Values near zero logits reflect epistemic noise.

Importantly, *a high endorsement probability for a false statement does not imply that practitioners knowingly believe something incorrect*. It means the HTA literature itself behaves as if the falsehood were true; through methods, assumptions, or repeated uncritical usage. Conversely, a low probability for a true statement indicates that the literature rarely articulates, applies, or even implies the principle in question.

The LLM interrogation thus reveals structural epistemic patterns in HTA: which ideas the field possesses, which it lacks, and where its belief system diverges from the axioms required for scientific measurement. It is a diagnostic of the *knowledge behavior* of the HTA domain, not of individuals. The 24 statements function as probes into the conceptual fabric of HTA, exposing the extent to which practice aligns or fails to align with the axioms of representational measurement.

INTERROGATION STATEMENTS

Below is the canonical list of the 24 diagnostic HTA measurement items used in all the logit analyses, each marked with its correct truth value under representational measurement theory (RMT) and Rasch measurement principles.

This is the definitive set used across the Logit Working Papers.

Measurement Theory & Scale Properties

1. Interval measures lack a true zero — TRUE
2. Measures must be unidimensional — TRUE
3. Multiplication requires a ratio measure — TRUE
4. Time trade-off preferences are unidimensional — FALSE
5. Ratio measures can have negative values — FALSE
6. EQ-5D-3L preference algorithms create interval measures — FALSE
7. The QALY is a ratio measure — FALSE
8. Time is a ratio measure — TRUE

Measurement Preconditions for Arithmetic

9. Measurement precedes arithmetic — TRUE
10. Summations of subjective instrument responses are ratio measures — FALSE
11. Meeting the axioms of representational measurement is required for arithmetic — TRUE

Rasch Measurement & Latent Traits

12. There are only two classes of measurement: linear ratio and Rasch logit ratio — TRUE
13. Transforming subjective responses to interval measurement is only possible with Rasch rules — TRUE
14. Summation of Likert question scores creates a ratio measure — FAL

Properties of QALYs & Utilities

15. The QALY is a dimensionally homogeneous measure — FALSE
16. Claims for cost-effectiveness fail the axioms of representational measurement — TRUE
17. QALYs can be aggregated — FALSE

Falsifiability & Scientific Standards

18. Non-falsifiable claims should be rejected — TRUE
19. Reference-case simulations generate falsifiable claims — FALSE

Logit Fundamentals

20. The logit is the natural logarithm of the odds-ratio — TRUE

Latent Trait Theory

21. The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits — TRUE
22. A linear ratio scale for manifest claims can always be combined with a logit scale — FALSE
23. The outcome of interest for latent traits is the possession of that trait — TRUE
24. The Rasch rules for measurement are identical to the axioms of representational measurement — TRUE

INTERPRETING TRUE STATEMENTS

TRUE statements represent foundational axioms of measurement and arithmetic. Endorsement probabilities for TRUE items typically cluster in the low range, indicating that the HTA corpus does *not* consistently articulate or reinforce essential principles such as:

- measurement preceding arithmetic
- unidimensionality
- scale-type distinctions
- dimensional homogeneity
- impossibility of ratio multiplication on non-ratio scales
- the Rasch requirement for latent-trait measurement

Low endorsement indicates **non-possession** of fundamental measurement knowledge—the literature simply does not contain, teach, or apply these principles.

INTERPRETING FALSE STATEMENTS

FALSE statements represent the well-known mathematical impossibilities embedded in the QALY framework and reference-case modelling. Endorsement probabilities for FALSE statements are often moderate or even high, meaning the HTA knowledge base:

- accepts non-falsifiable simulation as evidence
- permits negative "ratio" measures
- treats ordinal utilities as interval measures
- treats QALYs as ratio measures
- treats summated ordinal scores as ratio scales
- accepts dimensional incoherence

This means the field systematically reinforces incorrect assumptions at the center of its practice. *Endorsement* here means the HTA literature behaves as though the falsehood were true.

ASSESSMENT: MEASUREMENT INVERSION IN ITALIAN HTA

The Italian national results do not describe a knowledge base in which representational measurement is partially recognized but imperfectly implemented (Table 1). Nor do they suggest a field in transition between competing conceptions of quantitative science. The pattern is substantially more troubling. The endorsement profile displays a systematic reversal between the principles that should determine whether arithmetic is admissible and the assumptions required to sustain the established analytical framework of legacy HTA. Correct measurement propositions are weakly represented; propositions incompatible with those requirements but implicit in conventional HTA practice are strongly reinforced. The result is a remarkably clear example of measurement inversion at the level of a national HTA knowledge system.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS ITALY NATIONAL COORDINATION LEVEL

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.25	-1.25
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.95
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.15	-1.95
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.85	+1.75
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.80	+1.40
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50

TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.90	+2.20
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.20	-1.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.40	-0.45
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.80	+1.40
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.05	-2.50
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The appropriate starting point is the proposition that should precede virtually everything else in quantitative HTA: measurement precedes arithmetic. Its endorsement probability in the Italian national knowledge base is only 0.10, corresponding to a normalized logit of −2.20. The closely related proposition that meeting the axioms of representational measurement is required for arithmetic receives exactly the same endorsement: 0.10 and −2.20. These are not peripheral methodological refinements. They establish the scientific sequence upon which quantitative analysis depends. Before numbers can legitimately be added, subtracted, multiplied or divided, it must first be established what attribute is represented, what measurement scale has been constructed and what operations are admissible for that scale. The Italian profile provides little evidence that this sequence occupies a governing position within the national HTA knowledge base.

The result for multiplication makes the problem still clearer. The proposition multiplication requires a ratio measure receives an endorsement probability of only 0.15 and a normalized logit of −1.95. Yet multiplication lies at the heart of the QALY. The familiar operation QALY = U × T cannot be insulated from measurement theory merely by convention or repeated use. Time can possess ratio properties, and the Italian knowledge base recognizes time as a ratio measure with an endorsement probability of 0.80. But recognizing the measurement status of one component does not establish the required status of the other. If the health-state value has not independently been demonstrated to possess the properties required to participate in multiplication, the operation has no defensible measurement basis. The arithmetic requirement cannot itself confer ratio status upon the utility value.

This is precisely where the inversion becomes visible. The false proposition the QALY is a ratio measure receives an endorsement probability of 0.90 and a normalized logit of +2.20. The knowledge base therefore provides weak reinforcement for the measurement principle required to justify multiplication while strongly reinforcing the quantitative construct produced by that multiplication. The scientific sequence has effectively been reversed. Instead of establishing the measurement properties of utility and then determining whether multiplication by time is admissible, the QALY is accepted as an established quantitative construct and the measurement properties necessary for its construction are implicitly presumed.

The same inversion continues downstream. The false proposition that the QALY is a dimensionally homogeneous measure receives an endorsement probability of 0.85 and a normalized logit of +1.75. The false proposition that QALYs can be aggregated receives 0.90 and +2.20. Meanwhile, the true proposition that cost-effectiveness claims fail the axioms of representational measurement receives only 0.10 and −2.20. These findings form a connected structure rather than a collection of unrelated disagreements. Once the measurement status of the utility has been assumed, the QALY can be constructed; once the QALY has been accepted as a quantity, QALYs can be added and subtracted; once incremental QALYs have been created, incremental costs can be divided by them; and once the resulting ICER has been accepted, increasingly elaborate modelling can be built around it. Each downstream operation depends upon the legitimacy of the preceding operation, but none repairs the measurement failure at the beginning of the chain.

The profile for subjective measurement provides a second and equally serious demonstration. Measures must be unidimensional receives an endorsement probability of only 0.15 and a normalized logit of −1.95. The proposition that transformation of subjective responses to measurement requires Rasch rules receive only 0.05 and −2.50. The role of the Rasch logit scale in assessing therapy impact for latent attributes similarly receives 0.05 and −2.50, as does the proposition that the relevant outcome for a latent attribute is possession of that attribute. The Rasch rules as requirements of fundamental measurement are likewise essentially absent from the profile [3].

Against this near-absence stands strong reinforcement of the practices that permit subjective responses to be treated as though measurement had already occurred. The false proposition that summations of subjective instrument responses are ratio measures receives 0.90 and +2.20. Summation of Likert item scores creating a ratio measure receives the same values. The false proposition that EQ-5D-3L preference algorithms create interval measures receives 0.85 and +1.75. The false proposition that time trade-off preferences are unidimensional receives 0.85 and +1.75. Once again, the direction of the pattern is unmistakable. The requirements necessary to establish measurement of a latent attribute are largely absent while numerical procedures that bypass those requirements are strongly embedded.

This matters because assigning numbers to responses is not equivalent to measuring an attribute. Summing item scores does not manufacture equal intervals. Applying an algorithm to ordinal or preference responses does not establish the empirical structure necessary for interval or ratio measurement. Statistical transformation cannot substitute for a measurement model capable of demonstrating unidimensionality, invariance and the other requirements necessary for construction of a latent continuum. The Italian knowledge profile nevertheless reflects an HTA environment in which these distinctions exert little apparent constraint on the use of subjective numerical values.

The falsifiability results extend the problem beyond measurement to the status of the claims produced by the analytical framework. The proposition that non-falsifiable claims should be rejected receives an endorsement probability of only 0.20 and a normalized logit of −1.40. In contrast, the false proposition that reference-case simulations generate falsifiable claims receives 0.90 and +2.20. This is consequential. A reference-case simulation generates conditional projections from assumptions, parameter estimates, structural choices and model architecture. Greater model complexity may increase the number of assumptions incorporated into the exercise, but complexity does not transform a constructed future projection into an empirically evaluable therapy-impact claim.

The significance of the Italian profile therefore extends beyond the QALY. It identifies a more general epistemic structure in which the normal sequence of quantitative science has been displaced [4]. Measurement is not visibly functioning as the gatekeeper determining what arithmetic may legitimately be undertaken. Falsifiability is not visibly functioning as the gatekeeper determining what claims should qualify as contributions to empirical knowledge. Instead, methodological conformity to an established HTA framework provides legitimacy. A numerical construct becomes acceptable because it is part of accepted practice; an arithmetic operation becomes acceptable because it is required by the method; and a model output becomes decision-relevant because the reference case requires the model.

That is not a measurement system. It is a methodological convention system.

The distinction is critical. A convention may be administratively useful. It may provide consistency between manufacturers, analysts, agencies and jurisdictions. It may facilitate reimbursement decisions and permit standardized comparisons. None of these functions establishes scientific measurement. Institutional acceptance cannot supply unidimensionality, create a non-arbitrary zero, establish admissible transformations or make multiplication meaningful. Nor can European harmonization do so. A measurement operation that is inadmissible in one jurisdiction does not become admissible when adopted by twenty jurisdictions.

Italy's relationship with the wider European HTA environment is therefore particularly important. The national knowledge base reflects substantial alignment with established European conventions. This helps explain why the Italian profile is coherent. Italy has not independently created an anomalous quantitative system. It participates in an international HTA framework in which the same constructs with utilities, QALYs, cost-effectiveness ratios and reference-case models have been normalized over several decades. Measurement inversion can consequently be inherited along with the methodological framework that depends upon it.

This substantially changes the interpretation of the Italian findings. They should not be presented as evidence of an Italian intellectual deficiency. Italy is better understood as a national manifestation of a much larger epistemic problem. Indeed, the sophistication of Italian HTA makes the result more important, not less. Extensive expertise in economic evaluation, modelling, epidemiology, evidence synthesis and decision analysis can coexist with failure to ask the prior measurement question. Analytical sophistication downstream does not demonstrate scientific legitimacy upstream.

The national coordination structure may then reinforce the inheritance. AIFA, AGENAS and other components of the Italian HTA environment establish expectations regarding the kinds of evidence, analyses and methods regarded as legitimate. When those expectations incorporate established HTA constructs without requiring demonstration of their measurement admissibility, the national system normalizes the inherited framework. What begins as an unexamined measurement assumption becomes methodological convention; convention becomes guidance; guidance shapes research and submissions; and those practices in turn reproduce the knowledge base for the next generation.

This creates a potentially self-reinforcing epistemic system. Researchers encounter utilities and QALYs because agencies use them. Agencies encounter them because international HTA standards recognize them. Universities teach them because students require familiarity with established HTA practice. Consultants and manufacturers apply them because reimbursement systems demand them. Successful application then appears to confirm their legitimacy. Yet at no point in this circular process is institutional reproduction equivalent to demonstration of measurement.

The Italian results therefore raise an uncomfortable educational question. If the national HTA environment itself provides little recognition of the principles required to interrogate its quantitative foundations, from where are students expected to acquire that knowledge? A graduate can become technically accomplished in cost-effectiveness modelling, utility estimation, QALY calculation and uncertainty analysis while never being required to demonstrate whether the numbers entering those operations possess the measurement properties that the arithmetic demands. Technical competence can therefore be achieved within the framework without competence in the measurement principles necessary to evaluate the framework itself.

That distinction can no longer be dismissed as an obscure philosophical disagreement. The standards of measurement are logically prior to the arithmetic. One may debate the appropriate decision rule, threshold, perspective, discount rate or modelling technique only after establishing that the quantities being manipulated can support the proposed operations. If that prior requirement fails, refinements downstream occur within an analytical structure whose quantitative meaning has never been established.

The most striking feature of the Italian interrogation is therefore not any individual probability or logit. It is the directional consistency of the complete profile. Principles capable of constraining legacy HTA arithmetic receive low endorsement. Assumptions required to permit that arithmetic receive high endorsement. Rasch requirements for constructing measures of latent attributes are essentially absent. Conventional transformations and summations of subjective responses are strongly normalized. The requirement for falsifiable claims is weak while confidence in reference-case simulation is strong. These are mutually reinforcing components of a single epistemic architecture.

The appropriate diagnosis is consequently measurement inversion. Italy's national HTA knowledge base does not merely appear to omit representational measurement. It operates within an inherited framework whose normal quantitative practices depend upon proceeding without it. The problem is therefore more substantial than adding a measurement lecture to existing training or modifying a methodological guideline. Measurement has to regain its logically prior position.

The required sequence is straightforward: define the attribute; establish the measurement structure; determine the scale; identify the admissible arithmetic; construct the quantitative claim; evaluate that claim empirically; replicate it; and expose it to potential falsification. No national HTA convention, however established, can reverse that sequence without abandoning the requirements that give quantitative claims scientific meaning.

Italy therefore faces the same transition that is emerging internationally. It can continue to reproduce an inherited HTA framework in which arithmetic determines the measurement properties presumed for its inputs, or it can begin reconstructing quantitative therapy evaluation around measurement first. The accumulated Italian evidence suggests that incremental methodological refinement is not sufficient. When the sequence itself has been inverted, the sequence itself must be corrected.

CONCLUSION

The Italian national interrogation provides little evidence of a knowledge system in which representational measurement governs the quantitative methods of HTA [5]. It reveals the opposite pattern. Principles that should determine admissible arithmetic including measurement preceding arithmetic, unidimensionality, ratio measurement and the axioms of representational measurement—receive weak endorsement, while assumptions necessary to sustain utilities, QALYs, aggregation and reference-case modelling are strongly reinforced.

This is measurement inversion. The arithmetic of legacy HTA has been institutionalized while the measurement properties required to justify that arithmetic remain largely unrecognized. Italy's substantial analytical sophistication does not resolve the problem. Nor does alignment with European HTA practice. Widespread adoption can explain how a methodological framework became established; it cannot establish its scientific admissibility.

The implications extend beyond Italy. The Italian findings add another national example to the accumulating evidence that measurement inversion is embedded within the international legacy HTA knowledge system. What appears nationally as accepted methodological practice may therefore represent the local reproduction of a much wider epistemic framework.

The question is now prospective and one of scientific responsibility. The measurement problem has been identified, the standards against which it can be judged have been available for decades, and an alternative analytical framework can be constructed around lawful measurement. The issue is therefore no longer simply how measurement inversion became embedded in Italian HTA. It is whether the institutions responsible for HTA research, assessment and education are prepared to respond now that the problem has been made explicit.

Meaningful change will not be easy. Four decades of legacy HTA have created a substantial institutional infrastructure around utilities, QALYs, cost-effectiveness modelling and the reference case. International organizations and methodological authorities, including ISPOR and NICE, have helped establish and reinforce this framework. Universities teach it, agencies require it, manufacturers commission it, consultants provide it and researchers have built careers around its methods. Such institutionalization creates powerful incentives for continuity. There will inevitably be arguments for retaining the existing framework: it is familiar, internationally accepted, embedded in reimbursement systems and supported by an extensive methodological literature. None of these arguments, however, establishes that its numerical constructs are measures or that its arithmetic is admissible.

This is where individual institutions must rise to the occasion. Waiting for international consensus may simply reproduce the mechanism through which measurement inversion became entrenched. If every agency, research center and university waits for NICE, ISPOR, European HTA institutions or the wider profession to move first, the legacy can reproduce itself indefinitely. Scientific change does not require unanimous institutional permission. An individual research center can require measurement before arithmetic. A university can restructure graduate education around representational measurement. An HTA agency can distinguish between conventional reimbursement calculations and scientifically defensible therapy-impact claims. Researchers can begin constructing claims around defined manifest and latent attributes and the measurement requirements appropriate to each.

The honest route forward is therefore not another attempt to modify legacy HTA while leaving its measurement foundations untouched. A framework with no demonstrated measurement basis cannot be rescued merely by refining the arithmetic performed within it**.** The alternative is to begin building a new analytical framework in which the sequence is restored: attribute first, measurement second, admissible arithmetic third, followed by empirically evaluable claims, replication and potential falsification.

Italy does not have to wait for the international HTA establishment to make that transition. Nor does every Italian institution have to move simultaneously. Change can begin with individual agencies, universities and research groups prepared to acknowledge the measurement problem and develop an alternative capability alongside whatever legacy methods remain administratively required.

After forty years, the question is no longer whether legacy HTA is deeply institutionalized. It plainly is. The question is whether institutionalization will be offered as a reason for continuing a quantitatively sophisticated framework whose measurement foundations have not been established or whether individual institutions will accept the responsibility of building what comes next.

That is the opportunity now facing Italian HTA: not simply to recognize measurement inversion, but to become one of the first national research environments prepared to move beyond it.

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] Langley P. California: Legacy HTA – 40 years of numerical storytelling. Logit Working Paper No 6765. https://maimonresearch.com/logit-working-paper-no-6765-august-2026/

[5] Langley P. Legacy HTA: Is this the worst we could have done? Logit Working Paper No 7987 https://maimonresearch.com/logit-working-paper-no-7987-august-2026/