MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

THE MEASUREMENT FUTURE IN HEALTH TECHNOLGY ASSESSMENT

SPAIN: 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 1076 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

Background: *Spain provides a distinctive setting for examining the measurement foundations of health technology assessment (HTA). Its decentralized healthcare system combines regional assessment activity with national coordination through the Ministry of Health, the Spanish Network of Health Technology Assessment Agencies (RedETS), and pricing and reimbursement arrangements. This structure provides a natural test of whether institutional decentralization encourages independent examination of measurement assumptions or reproduces a common legacy HTA framework.*

Objective: *To assess whether the Spanish national coordinating HTA knowledge base recognizes and reinforces the principles of representational measurement required to establish measurement before arithmetic, or instead exhibits measurement inversion in which established quantitative operations precede demonstration of the measurement properties required to support them.*

Methods: *A large language model interrogation assessed the publicly represented Spanish HTA knowledge environment against 24 canonical TRUE/FALSE propositions addressing scale properties, representational measurement, admissible arithmetic, utilities, QALYs, Rasch measurement, latent attributes and falsifiability. Categorical endorsement probabilities were transformed to normalized logits ranging from −2.50 to +2.50. Probabilities represent conceptual reinforcement within the knowledge base, not statistical frequencies or beliefs of individual practitioners.*

Results: *The profile demonstrates systematic measurement inversion. Measurement preceding arithmetic was weakly endorsed (0.10; −2.20), as was the requirement that representational measurement axioms govern arithmetic (0.10; −2.20). Multiplication requiring ratio measurement received 0.20 (−1.40). Conversely, false propositions supporting established HTA arithmetic were strongly reinforced: the QALY as a ratio measure (0.85; +1.75), QALY aggregation (0.85; +1.75), and reference-case simulations as generating falsifiable claims (0.85; +1.75). Core Rasch propositions concerning latent measurement were essentially absent, typically receiving 0.05 (−2.50). The directional consistency indicates that principles capable of constraining legacy arithmetic are weakly represented while assumptions permitting that arithmetic are strongly reinforced.*

Conclusions: *Spain demonstrates that institutional decentralization does not necessarily produce epistemic independence. A common measurement failure can be reproduced across decentralized institutions through national and international methodological coordination. Yet decentralization also provides an opportunity for transition. Individual Spanish agencies, universities and research groups need not await national or European consensus to restore the required scientific sequence: attribute, measurement, scale, admissible arithmetic, empirically evaluable therapy-impact claims, replication and falsification.*

INTRODUCTION

Health technology assessment in Spain presents an unusual setting in which to examine the measurement foundations of contemporary HTA. Spain combines a highly decentralized healthcare system with substantial national coordination of assessment methods, evidentiary standards and policy expectations. Regional HTA agencies retain important roles within the autonomous communities, while the Ministry of Health, the Spanish Network of Health Technology Assessment Agencies (RedETS) and national pricing and reimbursement arrangements contribute to a shared methodological environment. The result is not a single centralized HTA authority, but a networked system in which methodological conventions can acquire national influence without being imposed through a single decision-making institution.

This structure raises an important scientific question. Does decentralization encourage independent examination of the measurement foundations of HTA, or do apparently independent institutions reproduce a common analytical framework whose underlying measurement assumptions remain unchallenged?

The question precedes debates over particular economic models, thresholds, perspectives, discount rates or methods for handling uncertainty. Quantitative science requires measurement before arithmetic. An attribute must first be defined; the properties required for its measurement must be established; the resulting scale must support the transformations and mathematical operations subsequently performed upon it. Numerical assignment alone does not establish measurement, and the ability to perform an arithmetic operation computationally does not establish that the operation has empirical meaning.

Legacy HTA reverses this scientific sequence when established analytical procedures determine the measurement properties that their inputs are implicitly assumed to possess. The QALY provides the clearest example. Multiplication of a health-state value by time requires that the participating quantities possess the measurement properties necessary for multiplication [1 2 3] . Subsequent aggregation of QALYs, subtraction to generate incremental QALYs and division of incremental costs by incremental QALYs introduce further arithmetic requirements. Decision modelling then combines these quantities through increasingly sophisticated sequences of calculation. None of these downstream operations can establish measurement properties that were absent at the beginning. Arithmetic cannot manufacture measurement.

Spain is particularly important because its institutional structure provides a natural test of whether methodological decentralization produces epistemic diversity. Regional autonomy might be expected to encourage alternative analytical approaches. Yet decentralization of institutions does not necessarily imply decentralization of scientific assumptions. If regional agencies operate within a common national and international methodological grammar, the same assumptions concerning utilities, QALYs, cost-effectiveness analysis and modelling may be reproduced throughout the system irrespective of organizational boundaries.

The national coordinating layer therefore assumes particular importance. RedETS explicitly promotes methodological consistency and shared standards across Spain's HTA agencies. The Ministry of Health provides a wider policy and coordinating environment, while pricing and

reimbursement processes reinforce expectations regarding the quantitative evidence considered relevant to national decision-making. These arrangements can provide considerable advantages in consistency, comparability and avoidance of unnecessary duplication. They can also, however, transmit methodological assumptions throughout a decentralized system.

The critical distinction is between methodological harmonization and scientific validation. Agreement among institutions about how an analysis should be performed does not establish that the quantities entering that analysis are measures. Repeated use of preference-based utilities does not establish their scale properties. Acceptance of QALYs does not establish that multiplication of utility by time is admissible. Standardization of cost-effectiveness modelling does not establish dimensional homogeneity, and agreement upon a reference case does not make its simulated claims empirically falsifiable. Harmonization can standardize a scientifically defensible framework, but it can equally standardize an inadmissible one.

The Spanish national HTA knowledge base is therefore interrogated here as a coordinating epistemic environment, not as a collection of individual opinions. The purpose is not to determine whether particular Spanish researchers, agency personnel or academics understand representational measurement. It is to determine whether the publicly represented HTA knowledge system recognizes and reinforces the measurement principles required to justify its quantitative operations. The interrogation employs 24 canonical propositions addressing representational measurement, scale properties, admissible arithmetic, utilities, QALYs, Rasch measurement, latent attributes and falsifiability.

The Spanish case consequently raises a question extending well beyond Spain. If a decentralized HTA system nevertheless reproduces a common failure to establish measurement before arithmetic, then organizational diversity alone offers no protection against measurement inversion. Indeed, coordination may strengthen the problem by transmitting the same analytical assumptions across otherwise independent institutions.

Spain therefore provides an important test of the epistemic architecture of legacy HTA. Has national coordination created consistency in analytical practice without first establishing the measurement foundations that would make that practice scientifically admissible**?** If so, the problem is not regional variation or insufficient methodological harmonization. It is the framework being harmonized.

1. 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 SPANISH NATIONAL HTA KNOWLEDGE BASE

The Spanish national HTA knowledge base is defined here as the publicly identifiable body of guidance, methodological conventions, institutional practices and analytical assumptions through which health technology assessment is coordinated, represented and reproduced across Spain. It should not be interpreted as the knowledge or beliefs of individual researchers, analysts or decision-makers. Nor does it imply that Spain possesses a single centralized HTA authority. The

object of interrogation is the collective epistemic environment within which Spanish HTA operates.

This distinction is particularly important because Spain has one of Europe's more decentralized healthcare systems. HTA activity is distributed across national and regional institutions associated with the autonomous communities. National coordination nevertheless provides a common methodological environment. The Ministry of Health contributes policy direction and coordination, while the Spanish Network of Health Technology Assessment Agencies (RedETS) brings together national and regional HTA bodies through shared assessments, methodological guidance and other mechanisms intended to promote consistency. Pricing and reimbursement processes associated with the Interministerial Commission on Medicine Prices (CIPM) provide an additional national environment within which particular forms of quantitative evidence acquire decision-making relevance.

The knowledge base therefore extends beyond individual guidance documents or the formal responsibilities of particular agencies. It encompasses the recurring quantitative constructs, methodological expectations and analytical practices through which Spanish HTA is conducted and communicated. These include economic evaluation, cost-effectiveness and cost-utility analysis, preference-based utilities, QALYs, patient-reported outcomes, evidence synthesis, decision modelling, sensitivity analysis and the treatment of uncertainty. Collectively, these practices reveal the measurement assumptions upon which the system operates, even where those assumptions are never explicitly articulated.

This is central to interpretation of the interrogation. A knowledge system can behave as though a proposition is true without formally declaring it to be true. Multiplication of a preference-based health-state value by time presumes that the quantities possess properties permitting multiplication. Aggregating QALYs presumes that QALYs support addition. Constructing incremental QALYs and dividing incremental costs by them similarly presupposes that the underlying quantities support subtraction and division. The analytical operations themselves therefore provide evidence of the measurement assumptions embedded within the knowledge base.

Spain must also be understood within the wider European and international HTA environment. The methods reproduced through Spanish coordination did not originate entirely within Spain. Utilities, QALYs, cost-effectiveness analysis and decision modelling were already established features of international HTA practice. National and regional alignment with these conventions can therefore transmit assumptions about measurement as readily as it transmits analytical procedures. Methodological harmonization cannot substitute for measurement demonstration**.**

The interrogation that follows should consequently be understood as an assessment of knowledge-system possession. It asks whether the Spanish coordinating environment recognizably possesses and reinforces the principles required by representational measurement and whether its established practices simultaneously reinforce propositions incompatible with those requirements. The resulting categorical endorsement probabilities are not survey estimates, measures of individual competence or assertions about what particular Spanish practitioners personally believe.

This national coordinating knowledge base provides the appropriate level at which to investigate measurement inversion in Spain. If the principles that should govern admissible arithmetic are weakly represented while the assumptions required to sustain legacy HTA arithmetic are strongly reinforced, decentralization does not remove the problem. It demonstrates instead that a common measurement failure can be transmitted through coordination across a decentralized HTA system.

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 an 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 — FALSE

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 SPANISH HTA

The Spanish national results reveal considerably more than an incomplete appreciation of measurement theory (Table 1). The pattern is systematic. Principles that should determine whether quantitative operations are scientifically admissible receive weak endorsement, while assumptions necessary to sustain the established arithmetic of legacy HTA receive strong reinforcement. The result is a clear example of measurement inversion within the national coordinating knowledge environment: arithmetic has acquired methodological authority without prior demonstration that the quantities entering that arithmetic possess the measurement properties the operations require.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS SPAIN NATIONAL COORDINATION IN HTA

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	-1.10
MEASURES MUST BE UNIDIMENSIONAL	1	0.20	-1.40
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.20	-1.40
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.80	+1.40
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.85	+1.75
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.80	+1.40
THE QALY IS A RATIO MEASURE	0	0.85	+1.75
TIME IS A RATIO MEASURE	1	0.85	+1.75
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20

SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.85	+1.75
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	0.85	+1.75
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.80	+1.40
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.10	-2.20
QALYS CAN BE AGGREGATED	0	0.85	+1.75
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.20	-1.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
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	0.75	+1.10
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 measurement precedes arithmetic. Its endorsement probability is only 0.10, corresponding to a normalized logit of −2.20. The

proposition that meeting the axioms of representational measurement is required for arithmetic receives exactly the same values. These are not specialist propositions peripheral to HTA methodology. They express the logically prior requirement of quantitative science. Before addition, subtraction, multiplication or division can be interpreted empirically, the attribute must be defined, measurement established, the scale properties demonstrated and the arithmetic shown to be admissible for that scale. The Spanish profile provides little evidence that this sequence functions as a governing constraint upon contemporary HTA.

The result for multiplication makes the problem particularly clear. The proposition that multiplication requires a ratio measure receives an endorsement probability of only 0.20 and a normalized logit of −1.40. Yet multiplication is fundamental to construction of the QALY. Spain's knowledge base strongly recognizes time as a ratio measure, with an endorsement probability of 0.85 and logit of +1.75. Recognition of the measurement properties of time, however, cannot confer equivalent properties upon the health-state value by which time is multiplied. Those properties must be demonstrated independently before the operation (U \times T) can be given empirical meaning.

The inversion becomes apparent when this result is compared with the false proposition that the QALY is a ratio measure. That proposition receives an endorsement probability of 0.85 and a normalized logit of +1.75. The measurement requirement necessary to justify the arithmetic is therefore weakly represented while the product of that arithmetic is strongly normalized as a quantitative construct. Instead of measurement properties determining whether the QALY operation is admissible, acceptance of the QALY appears to presume the properties required to make the operation possible.

The failure then propagates downstream. The false proposition that the QALY is dimensionally homogeneous receives 0.80 and +1.40, while the false proposition that QALYs can be aggregated receives 0.85 and +1.75. Against these results, the true proposition that cost-effectiveness claims fail the axioms of representational measurement receives only 0.10 and −2.20. The structure is internally consistent. Once utility is accepted as though it were an appropriate quantity, multiplication by time becomes conventional; once QALYs are accepted as quantities, they can be aggregated and differenced; once incremental QALYs are accepted, incremental costs can be divided by them to produce the ICER. Additional modelling, sensitivity analysis and probabilistic analysis can then be built around the resulting construct. None of these later operations repairs a measurement failure inherited from the first.

The treatment of subjective attributes reveals the same inversion. The requirement that measures be unidimensional receives an endorsement probability of only 0.20 and a normalized logit of −1.40. More strikingly, propositions concerning Rasch measurement are essentially absent. Transformation of subjective responses to measurement under Rasch rules receives 0.05 and −2.50; the Rasch logit ratio scale as the basis for assessing therapy impact for latent attributes receives 0.05 and −2.50; possession of the latent attribute as the outcome of interest receives 0.05 and −2.50; and the correspondence between Rasch requirements and the axioms of measurement similarly receives 0.05 and −2.50.

Against this near-absence of measurement requirements stands substantial reinforcement of numerical procedures that bypass them. The false proposition that summated subjective instrument responses are ratio measures receives 0.85 and +1.75. The false proposition that summation of Likert scores creates a ratio measure receives the same values. The proposition that EQ-5D-3L preference algorithms create interval measures, also false under the framework applied here, receives 0.80 and +1.40. Time trade-off preferences being unidimensional receives 0.80 and +1.40. The pattern again runs in one direction: the requirements for constructing a measure of a latent attribute are weakly represented, while procedures that permit numerical scores to be treated as measures are strongly reinforced.

This distinction cannot be resolved by greater statistical sophistication. Assigning numbers to categories does not establish measurement. Summing responses does not create equal units. Applying an algorithm does not demonstrate unidimensionality, invariance or the required scale structure. Nor does repeated use of a numerical score establish that ratios or differences between values possess empirical meaning. Measurement properties must be demonstrated before the arithmetic begins; they cannot be inferred retrospectively from the fact that the arithmetic has become conventional.

The results for falsifiability extend the problem beyond measurement. 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 0.85 and +1.75. Reference-case models can be useful conditional decision exercises, but their complexity cannot transform assumptions about future events into independently testable therapy-impact claims. A model can calculate precisely what follows from its assumptions without demonstrating that the resulting projection represents an empirically discoverable quantity.

Taken together, the Spanish results describe an epistemic architecture in which methodological convention has displaced measurement as the gatekeeper of quantitative legitimacy. Utilities are accepted because they are established HTA inputs; QALYs are accepted because they are established HTA outcomes; cost-effectiveness ratios are accepted because they are required by economic evaluation; and models are accepted because modelling has become part of the expected analytical framework. The legitimacy of each component is reinforced by the legitimacy attributed to the system as a whole.

The distinctive importance of Spain, however, lies in how this framework is reproduced. Spain does not possess a single centralized HTA authority capable of imposing every element of the analytical framework upon the entire country. Its HTA system is decentralized, while RedETS and other national coordinating arrangements provide methodological consistency across regional structures. The finding of measurement inversion therefore cannot readily be attributed to the intellectual preferences of a single dominant national agency. Instead, it suggests that a shared epistemic grammar can operate across institutionally separate organizations.

This is an important distinction. Institutional decentralization has not produced epistemic decentralization. Regional organizations may differ in responsibilities, technologies assessed,

procedures and relationships with decision-makers while continuing to employ the same underlying assumptions about utilities, QALYs, economic evaluation and quantitative evidence. Coordination can therefore reproduce measurement inversion without requiring centralized analytical control.

Spain's relationship with wider European HTA reinforces this process. The analytical framework being coordinated nationally is itself substantially inherited from international HTA. Spain did not need independently to establish the legitimacy of the QALY, preference-based utility measurement or reference-case modelling because these had already acquired professional authority internationally. European and international methodological alignment can consequently transmit measurement assumptions together with methods. A practice accepted across multiple jurisdictions may appear increasingly authoritative precisely because each jurisdiction observes its acceptance elsewhere.

But replication of a convention is not replication of a scientific result. Twenty agencies employing the same inadmissible operation do not transform it into an admissible operation. European harmonization cannot establish unidimensionality, create a non-arbitrary zero, confer ratio properties upon an ordinal score or make an otherwise non-falsifiable claim falsifiable. Consensus can establish a standard of practice; it cannot establish a scale of measurement.

There is, however, another side to Spain's decentralized structure, and it may prove more important for the future than for explaining the past. Spain has an unusual opportunity to respond to the measurement failure identified here**.** The institutional degrees of freedom inherent in a decentralized system mean that scientific transition need not await simultaneous reform of the entire national HTA infrastructure. A regional HTA agency, university research center, methodological group or collaborating network could begin independently to require measurement before arithmetic and develop an alternative analytical framework around that principle.

This gives Spain a potential advantage that more centralized European HTA systems may not possess. Where methodological authority is concentrated within a dominant national institution, fundamental change may depend heavily upon that institution first accepting the need for reform. Spain offers multiple points of entry. One component of the system can innovate without requiring every other component to move at the same time. Alternative approaches can be developed, applied, evaluated and demonstrated within the existing decentralized structure. Successful methodological innovation could then move horizontally through networks such as RedETS rather than waiting for change to descend vertically from a single national authority.

The same decentralization that has allowed a common legacy framework to be reproduced could therefore become the mechanism through which that framework is challenged. This is potentially a unique opportunity within European HTA. Spain could treat decentralization not simply as an administrative feature but as a scientific resource: a structure capable of supporting methodological experimentation and permitting individual institutions to rise to the challenge without awaiting international consensus.

That opportunity also creates responsibility. Once the measurement problem has been made explicit, continued conformity cannot be defended solely by pointing to European practice,

international guidelines or four decades of established HTA methodology. Those facts explain why the framework became institutionalized. They do not establish its scientific validity. Nor does the continuing authority of organizations that promote conventional HTA remove the responsibility of individual institutions to determine whether the quantitative methods they employ satisfy the requirements of measurement.

The most important feature of the Spanish interrogation is therefore not any single probability or normalized logit. It is the directional consistency of the complete profile. Measurement principles capable of constraining legacy arithmetic receive weak endorsement. Assumptions necessary to sustain that arithmetic receive strong endorsement. Requirements for measurement of latent attributes are virtually absent, while numerical transformations that bypass those requirements are normalized. Falsifiability receives limited recognition while confidence in reference-case simulation remains strong. These findings are mutually reinforcing components of a common structure.

The appropriate diagnosis is measurement inversion. The appropriate response cannot therefore be limited to refining models, improving uncertainty analysis or adding further methodological sophistication to the existing framework. If measurement has been placed after arithmetic, the scientific sequence itself must be restored.

That sequence is straightforward: define the attribute; establish its measurement requirements; construct the appropriate measure; determine the scale and admissible transformations; identify the arithmetic that the measure supports; formulate the therapy-impact claim; evaluate that claim empirically; replicate it; and expose it to potential falsification.

Spain consequently faces both a failure and an opportunity. Its national coordinating knowledge base reflects the measurement inversion inherited by contemporary HTA, but its decentralized institutional structure provides unusual freedom to begin moving beyond it. Spain need not wait for Europe to change before one Spanish institution changes**.** A regional agency, research group or university willing to place measurement before arithmetic could provide the demonstration that a different analytical framework is possible.

In that respect, Spain may be better positioned than many European systems to make the transition. The question is whether decentralization will continue merely to reproduce a common inherited framework or whether its degrees of freedom will be used to create what comes next.

Spain's decentralized structure provides an unusual opportunity to respond. The transition need not depend upon simultaneous reform of the entire national HTA system or agreement across European HTA institutions. Individual regional agencies, universities and research groups have sufficient institutional freedom to begin with the fundamental requirement: measurement must precede arithmetic. This does not require the creation of another elaborate methodological structure. It requires quantitative claims to meet the established requirements of measurement theory before arithmetic is applied.

For manifest attributes, this means establishing the properties required for linear ratio measurement before undertaking arithmetic that depends upon those properties. For latent attributes, it means

constructing measures capable of satisfying the requirements of conjoint measurement rather than treating assigned, summed or algorithmically transformed scores as though numerical representation itself established measurement. In both cases, the principle is the same: the attribute must be defined, measurement established and the admissible arithmetic determined before a quantitative therapy-impact claim is constructed.

Spain may therefore possess an unusual degree of freedom within European HTA. Its decentralized structure provides multiple points at which the scientific sequence can be restored without waiting for a single national authority, European institution or international professional organization to authorize the transition. No new scientific standard has to be invented. The immediate requirement is simply to meet the standards of measurement that quantitative science already requires.

CONCLUSION

The Spanish interrogation identifies a clear and consequential pattern. The national coordinating HTA knowledge base provides little evidence that the requirements of representational measurement operate as prior conditions for quantitative analysis. Measurement preceding arithmetic, the requirement for ratio measurement before multiplication, unidimensionality, the axioms governing admissible arithmetic and the measurement requirements for latent attributes are weakly represented. At the same time, utilities, QALYs, aggregation, cost-effectiveness analysis and reference-case modelling are strongly embedded within accepted HTA practice.

The diagnosis is measurement inversion. Arithmetic has been institutionalized before the measurement properties required to give that arithmetic empirical meaning have been demonstrated [4] [5]. The problem is therefore not that Spanish HTA lacks analytical sophistication. It possesses substantial methodological capability. The problem is that sophistication has developed within an inherited framework in which measurement has not been allowed to perform its logically prior scientific role.

Spain did not create this problem. Its HTA system has inherited much of the analytical architecture of international and European HTA. Utilities, QALYs, cost-effectiveness modelling and associated decision frameworks arrived with considerable professional authority already attached to them. National coordination through RedETS and the wider Spanish HTA environment has helped reproduce these conventions across a decentralized system. But international acceptance cannot establish measurement. European harmonization cannot make inadmissible arithmetic admissible. Repetition of a quantitative convention across countries cannot supply measurement properties that were never demonstrated in the first place. The significance of Spain lies in what happens next.

Spain has an opportunity that may be unusually important within European HTA. Its decentralized institutional structure provides degrees of freedom that more centralized systems may lack. Change does not require every regional agency, university, research center and national institution to move simultaneously. Nor does Spain have to wait for a European consensus, a revised international reference case or permission from established methodological authorities. An individual Spanish institution can begin by requiring something much more fundamental: that measurement be established before arithmetic is undertaken.

The requirement is neither radical nor new. It is a return to the elementary standards of quantitative science. Define the attribute. Determine whether it is manifest or latent. Establish the appropriate measurement structure. Demonstrate the properties of the resulting scale. Apply only those mathematical operations admissible for that scale. Construct therapy-impact claims that can be evaluated empirically, replicated and exposed to potential falsification. Nothing more elaborate is required to begin the transition.

This also changes the responsibility facing Spanish HTA institutions. The argument that legacy methods are internationally accepted can explain their historical adoption, but it cannot determine their future use once their measurement foundations have been challenged. Universities have an opportunity to ensure that the next generation of HTA researchers understands measurement before learning the arithmetic built upon it. Research groups can require measurement properties to be demonstrated rather than assumed. Regional agencies can ask whether the quantities presented to them are actually measures before accepting calculations based upon them.

Spain therefore need not approach the present findings defensively. Decentralization can become an advantage. A system containing multiple centers of expertise and decision-making also contains multiple places from which scientific reconstruction can begin.

The choice is consequently not between preserving contemporary HTA unchanged and abandoning quantitative evaluation. It is between continuing a legacy framework whose measurement foundations have not been established and building quantitative evaluation upon foundations that can be established.

Spain is particularly well placed to make that choice. It can move from coordination around inherited methodological conventions toward coordination around the requirements of measurement itself. It can place the attribute before the instrument, measurement before arithmetic, and empirical evaluation before simulated certainty.

Spain does not need to invent a new standard of quantitative science. It needs only to insist that HTA meet the standards that quantitative science already requires. That is not a retreat from sophisticated HTA. It is the opportunity to build something better**.**

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/