MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

ISPOR GLOBAL NETWORK: THE ENDORSEMENT OF MEASUREMENT FAILURE - LATIN AMERICA CONSORTIUM

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

LOGIT WORKING PAPER No 4018 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF MEASUREMENT INVERSION FAILURE

ABSTRACT

This study interrogates the health technology assessment (HTA) knowledge base associated with the ISPOR Latin America Consortium to determine whether it recognizes the requirements of representational measurement or reproduces the measurement inversion characteristic of contemporary health economics and outcomes research (HEOR). The central criterion is that measurement must precede arithmetic.

A large language model interrogation was undertaken using 24 canonical true and false statements derived from representational measurement theory and Rasch measurement principles. Statements address unidimensionality, scale properties, ratio measurement, admissible arithmetic, dimensional homogeneity, utilities, QALYs, manifest and latent attributes, Rasch measurement and falsifiability. Categorical endorsement probabilities were assigned to each statement and transformed to normalized logits to characterize the conceptual structure of the regional knowledge base.

The interrogation identified a consistent pattern of measurement inversion. Foundational propositions received weak endorsement, including measurement preceding arithmetic, multiplication requiring ratio measurement, the requirement to satisfy representational measurement axioms, and the role of Rasch measurement for latent attributes. Conversely, false propositions embedded in conventional HEOR received strong endorsement. These included treating preference algorithms and summed subjective scores as measures, accepting QALYs as ratio and dimensionally homogeneous quantities, aggregating QALYs, and treating reference-case simulations as generating falsifiable claims. The findings indicate an environment in which arithmetic and modelling are accepted without first establishing the measurement properties required to support them.

The findings should not be interpreted as a specifically Latin American failure. They reflect transmission of an established international HEOR framework in which utilities, QALYs, cost-effectiveness analysis, mapping and simulation modelling have been institutionalized without adequate attention to their measurement foundations. This is particularly important as HTA capacity expands across Latin America because measurement inversion may become embedded in guidelines, curricula, professional training and reimbursement systems. The required response is reconstruction around lawful measurement. Therapy-impact assessment should begin with definition of the attribute and establishment of the appropriate measurement system. Manifest

attributes require lawful linear ratio measures and latent attributes require invariant Rasch logit ratio measurement. Arithmetic should follow measurement, with claims prospectively specified, empirically evaluated and open to falsification. The transition is therefore not away from quantitative HTA, but toward a scientifically defensible quantitative framework.

INTRODUCTION: ISPOR GLOBAL NETWORK LATIN AMERICA CONSORTIUM

The ISPOR Global Network provides an international structure through which the Society promotes health economics and outcomes research (HEOR), pharmacoeconomics, health technology assessment (HTA), economic evaluation and their application to healthcare decision-making. Regional networks and country chapters provide an important mechanism for extending the ISPOR methodological framework into local academic, professional and policy environments.

ISPOR formally identifies a Latin America Consortium within its Global Groups structure, while its Regional Chapter Directory identifies country and regional chapters across Latin America. For the purposes of the present review, the Latin America knowledge base is defined by this regional ISPOR structure and comprises Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, Mexico, Panama, Peru and Uruguay. These countries encompass markedly different stages of development in health economics, HTA, reimbursement systems and formal economic evaluation, but are connected through the broader ISPOR regional HEOR framework.

These country chapters and associated regional activities provide a structure through which ISPOR principles, educational activities and HEOR methodologies can be disseminated and applied. Their importance extends well beyond professional networking. As HTA and formal economic evaluation continue to develop across Latin America, the methodological framework promoted through ISPOR can influence academic research, professional education, national HTA systems, reimbursement assessment, formulary decision making and healthcare policy. This is particularly important in a region containing both established HTA institutions and countries currently developing or expanding formal HTA capacity. The Latin America Consortium therefore provides an important regional knowledge base for assessing whether contemporary HEOR recognizes the requirements of lawful measurement before undertaking quantitative evaluations of therapy impact.

The objective of this study is to determine whether the ISPOR Latin America knowledge base endorses the requirements for lawful measurement in the quantitative evaluation of therapy impact or, alternatively, endorses measurement inversion. Using 24 diagnostic statements, the interrogation assesses recognition of representational measurement, scale properties, admissible arithmetic, dimensional homogeneity, manifest and latent attributes, Rasch measurement, utilities, QALYs and falsifiability. The fundamental test is whether measurement is required to precede arithmetic. Where this requirement is absent and numerical operations are undertaken without first establishing lawful measures, the consequence is measurement inversion and, for quantitative claims of therapy impact, measurement failure.

The findings demonstrate a consistent endorsement of measurement inversion and, therefore, measurement failure. Foundational requirements for lawful measurement receive low endorsement, including measurement preceding arithmetic, ratio measurement for multiplication and the axioms of representational measurement. In contrast, false statements supporting conventional HEOR practices, including the ratio status, dimensional homogeneity and aggregation of QALYs, receive strong endorsement. The Latin America knowledge base therefore reproduces the measurement-failure pattern previously identified within the parent ISPOR framework.

The starting point is simple and inescapable: *measurement precedes arithmetic*. This principle is not a methodological preference but a logical necessity. One cannot multiply what one has not measured, cannot sum what has no dimensional homogeneity, cannot compare ratios when no ratio scale exists. When HTA multiplies time by utilities to generate QALYs, it is performing arithmetic with numbers that cannot support the operation. When HTA divides cost by QALYs, it is constructing a ratio from quantities that have no ratio properties. When HTA aggregates QALYs across individuals or conditions, it is combining values that do not share a common scale. These practices are not merely suboptimal; they are mathematically impossible.

The modern articulation of this principle can be traced to Stevens' seminal 1946 paper, which introduced the typology of nominal, ordinal, interval, and ratio scales [1]. Stevens made explicit what physicists, engineers, and psychologists already understood: different kinds of numbers permit different kinds of arithmetic. Ordinal scales allow ranking but not addition; interval scales permit addition and subtraction but not multiplication; ratio scales alone support multiplication, division, and the construction of meaningful ratios. Utilities derived from multiattribute preference exercises, such as EQ-5D or HUI, are ordinal preference scores; they do not satisfy the axioms of interval measurement, much less ratio measurement. Yet HTA has, for forty years, treated these utilities as if they were ratio quantities, multiplying them by time to create QALYs and inserting them into models without the slightest recognition that scale properties matter. Stevens' paper should have blocked the development of QALYs and cost-utility analysis entirely. Instead, it was ignored.

The foundational theory that establishes *when* and *whether* a set of numbers can be interpreted as measurements came with the publication of Krantz, Luce, Suppes, and Tversky's *Foundations of Measurement* (1971) [2] . Representational Measurement Theory (RMT) formalized the axioms under which empirical attributes can be mapped to numbers in a way that preserves structure. Measurement, in this framework, is not an act of assigning numbers for convenience, it is the discovery of a lawful relationship between empirical relations and numerical relations. The axioms of additive conjoint measurement, homogeneity, order, and invariance specify exactly when interval scales exist. RMT demonstrated once and for all that measurement is not optional and not a matter of taste: either the axioms hold and measurement is possible, or the axioms fail and measurement is impossible. Every major construct in HTA, utilities, QALYs, DALYs, ICERs, incremental ratios, preference weights, health-state indices, fails these axioms. They lack unidimensionality; they violate independence; they depend on aggregation of heterogeneous attributes; they collapse under the requirements of additive conjoint measurement. Yet HTA proceeded, decade after decade, without any engagement with these axioms, as if the field had collectively decided that measurement theory applied everywhere except in the evaluation of therapies.

Whereas representational measurement theory articulates the axioms for interval measurement, Georg Rasch's 1960 model provides the only scientific method for transforming ordered categorical responses into interval measures for latent traits [3]. Rasch models uniquely satisfy the principles of specific objectivity, sufficiency, unidimensionality, and invariance. For any construct such as pain, fatigue, depression, mobility, or need, Rasch analysis is the only legitimate means of producing an interval scale from ordinal item responses. Rasch measurement is not an alternative to RMT; it is its operational instantiation. The equivalence of Rasch's axioms and the axioms of representational measurement was demonstrated by Wright, Andrich and others as early as the 1970s. In the latent-trait domain, the very domain where HTA claims to operate; Rasch is the only game in town [4].

Yet Rasch is effectively absent from all HTA guidelines, including NICE, PBAC, CADTH, ICER, SMC, and PHARMAC. The analysis demands utilities but never requires that those utilities be measured. They rely on multiattribute ordinal classifications but never understand that those constructs be calibrated on interval or ratio scales. They mandate cost-utility analysis but never justify the arithmetic. They demand modelled QALYs but never interrogate their dimensional properties. These guidelines do not misunderstand Rasch; they do not know it exists. The axioms that define measurement and the model that makes latent trait measurement possible are invisible to the authors of global HTA rules. The field has evolved without the science that measurement demands.

How did HTA and then ISPOR miss the bus so thoroughly? The answer lies in its historical origins. In the late 1970s and early 1980s, HTA emerged not from measurement science but from welfare economics, decision theory, and administrative pressure to control drug budgets. Its core concern was *valuing health states*, not *measuring health*. This move, quiet, subtle, but devastating, shifted the field away from the scientific question "What is the empirical structure of the construct we intend to measure?" and toward the administrative question "How do we elicit a preference weight that we can multiply by time?" The preference-elicitation projects of that era (SG, TTO, VAS) were rationalized as measurement techniques, but they never satisfied measurement axioms. Ordinal preferences were dressed up as quasi-cardinal indices; valuation tasks were misinterpreted as psychometrics; analyst convenience replaced measurement theory. The HTA community built an entire belief system around the illusion that valuing health is equivalent to measuring health. It is not.

The endurance of this belief system, forty years strong and globally uniform, is not evidence of validity but evidence of institutionalized error. HTA has operated under conditions of what can only be described as *structural epistemic closure*: a system that has never questioned its constructs because it never learned the language required to ask the questions. Representational measurement theory is not taught in graduate HTA programs; Rasch modelling is not part of guideline development; dimensional analysis is not part of methodological review. The field has been insulated from correction because its conceptual foundations were never laid. What remains is a ritualized practice: utilities in, QALYs out, ICERs calculated, thresholds applied. The arithmetic continues because everyone assumes someone else validated the numbers.

This Logit Working Paper series exposes, through probabilistic and logit-based interrogations of AI large language national knowledge bases, the scale of this failure. The results display a global pattern: true statements reflecting the axioms of measurement receive weak endorsement; false statements reflecting the HTA belief system receive moderate or strong reinforcement. This is not disagreement. It is non-possession. It shows that HTA, worldwide, has developed as a quantitative discipline without quantitative foundations; a confused exercise in numerical storytelling.

The conclusion is unavoidable: HTA does not need incremental reform; it needs a scientific revolution. Measurement must precede arithmetic. Representational axioms must precede valuation rituals. Rasch measurement must replace ordinal summation and utility algorithms. Value claims must be falsifiable, protocol-driven, and measurable; rather than simulated, aggregated, and numerically embellished.

The global system of non-measurement is now visible. The task ahead is to replace it with science.

Paul C Langley, Ph.D

Email: langleylapaloma@gmail.com

DISCLAIMER

This analysis is generated through the structured interrogation of a large language model (LLM) applied to a defined documentary corpus and is intended solely to characterize patterns within an aggregated knowledge environment. It does identify, assess, or attribute beliefs, intentions, competencies, or actions to any named individual, faculty member, student, administrator, institution, or organization. The results do not constitute factual findings about specific persons or programs, nor should they be interpreted as claims regarding professional conduct, educational quality, or compliance with regulatory or accreditation standards. All probabilities and logit values reflect model-based inferences about the presence or absence of concepts within a bounded textual ecosystem, not judgments about real-world actors. The analysis is exploratory, interpretive, and methodological in nature, offered for scholarly discussion of epistemic structures rather than evaluative or legal purposes. Any resemblance to particular institutions or practices is contextual and non-attributive, and no adverse implication should be inferred.

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 ISPOR LATIN AMERICA CONSORTIUM HTA KNOWLEDGE BASE

The ISPOR Latin America Consortium HTA knowledge base is not confined to material formally published by ISPOR or to the activities of individual country chapters. It is better understood as the interconnected body of knowledge through which contemporary health economics and outcomes research (HEOR), pharmacoeconomics, health technology assessment (HTA) and economic evaluation are taught, promoted, published and applied across Latin America. ISPOR provides an important international methodological framework, while country chapters, universities, research centers, government agencies, HTA organizations, healthcare systems and individual investigators provide the institutional environments through which that framework is interpreted, transmitted and implemented.

The knowledge base therefore encompasses ISPOR Latin America Consortium and country-chapter activities; regional conferences, workshops and educational programs; publications by chapter members and associated researchers; university teaching and research in health economics, pharmacoeconomics and HTA; national HTA agencies and methodological guidelines; reimbursement and formulary assessment frameworks; and published economic evaluations undertaken throughout the region. Particular importance attaches to material addressing cost-effectiveness analysis, utilities, QALYs, preference-based instruments, mapping, patient-reported outcomes, decision-analytic modelling, sensitivity analysis, budget impact and value assessment. Collectively, these sources define the quantitative assumptions embedded within the regional HEOR environment.

The Latin America knowledge base is particularly important because the region encompasses healthcare systems at different stages of HTA institutionalization. Some countries have well-developed structures for technology assessment, economic evaluation, pricing and reimbursement, while others are developing or expanding formal HTA capabilities. Despite these institutional differences, a recognizable methodological language extends across the region. Cost-effectiveness analysis, preference-based health-state assessment, QALYs, ICERs, modelling and sensitivity analysis increasingly provide the framework through which quantitative value claims are expressed. The ISPOR Latin America Consortium provides one mechanism through which these internationally established methods are communicated across otherwise diverse healthcare systems.

This broader definition is essential because methodological transmission does not occur solely through formal ISPOR publications. A country chapter may itself produce relatively little methodological guidance while its members teach pharmacoeconomics, conduct economic evaluations, participate in HTA committees, advise reimbursement authorities and train subsequent generations of practitioners. Universities, government agencies and research organizations may similarly employ methods consistent with the international HEOR framework without formally attributing those practices to ISPOR. The relevant knowledge base is therefore the publicly accessible intellectual environment associated with the Latin America Consortium rather than a narrowly bounded collection of official ISPOR documents.

For interrogation purposes, the critical question is what this knowledge environment indicates about the status of measurement. Evidence is sought for recognition that quantitative evaluation must begin by defining the attribute to be measured and establishing the properties of the resulting scale before arithmetic is undertaken. This includes unidimensionality, scale type, admissible transformations, dimensional homogeneity, manifest and latent attributes, ratio measurement and Rasch measurement for latent attributes. These requirements are contrasted with the acceptance of utilities, QALYs, composite scores, mapping, aggregation and simulation modelling as quantitative inputs and outputs.

The interrogation does not assume methodological uniformity across Latin America, nor that every publication associated with a country chapter represents an official ISPOR position. Its purpose is to characterize the dominant publicly accessible knowledge environment associated with the ISPOR Latin America Consortium. The fundamental question is whether that environment places measurement before arithmetic or reproduces a framework in which numerical operations are accepted before lawful measurement has been established. The latter constitutes measurement inversion and, when used to support quantitative claims of therapy impact, measurement failure.

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

AI LARGE LANGUAGE MODEL STATEMENTS: TRUE OR FALSE

Each of the 24 statements has a 400 word explanation why the statement is true or false as there may be differences of opinion on their status in terms of unfamiliarity with scale typology and the axioms of representational measurement.

The link to these explanations is: https://maimonresearch.com/ai-llm-true-or-false/

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.

2. SUMMARY OF FINDINGS FOR TRUE AND FALSE ENDORSEMENTS: THE LATIN AMERICAN CONSORTIUM HTA KNOWLEDGE BASE

Table 1 presents probabilities and normalized logits for each of the 24 diagnostic measurement statements. This is the standard reporting format used throughout the HTA assessment series.

It is essential to understand how to interpret these results.

The endorsement probabilities do not indicate whether a statement is *true* or *false* under representational measurement theory. Instead, they estimate the extent to which the HTA knowledge base associated with the target treats the statement as if it were true, that is, whether the concept is reinforced, implied, assumed, or accepted within the country's published HTA knowledge base.

The logits provide a continuous, symmetric scale, ranging from +2.50 to –2.50, that quantifies the degree of this endorsement. the logits, of course link to the probabilities (p) as the logit is the natural logarithm of the odds ratio; logit = ln[p/1-p].Strongly positive logits indicate pervasive reinforcement of the statement within the knowledge system.

- Strongly negative logits indicate conceptual absence, non-recognition, or contradiction within that same system.
- Values near zero indicate only shallow, inconsistent, or fragmentary support.

Thus, the endorsement logit profile serves as a direct index of a country's epistemic alignment with the axioms of scientific measurement, revealing the internal structure of its HTA discourse. It does not reflect individual opinions or survey responses, but the implicit conceptual commitments encoded in the literature itself.

ASSESSMENT OF THE ISPOR LATIN AMERICAN KNOWLEDGE BASE CONSORTIUM

The 24 canonical statements were applied to the publicly accessible ISPOR-related Latin America consortium knowledge base. The interrogation asks whether this knowledge base would endorse each statement, with endorsement represented by a categorical probability and transformed to the standardized normalized logit range of ±2.50. The response column identifies the stipulated measurement status of each statement: 1 = true and 0 = false. The critical issue is therefore not whether a statement is endorsed, but whether the knowledge base endorses statements consistent with lawful measurement or statements representing measurement inversion. The results are presented in Table 1.

TABLE 1: ITEM STATEMENT RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS ISPOR LATIN AMERICA KNOWLEDGE BASE

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.20	-1.40
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.15	-1.75

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	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.90	+2.20
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.15	+1.75
SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.15	-1.75
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	0.90	+2.20
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.65	+0.60
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.85	+1.75
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.65	+0.60
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	0	0.60	+0.40

BE COMBINED WITH A LOGIT SCALE			
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.20	-1.40
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The interrogation presents a strikingly consistent measurement-inversion profile for the ISPOR-related Latin America knowledge base. The dominant regional framework is not one in which attributes are first defined, measurement properties established and permissible arithmetic subsequently determined. Instead, the numerical methods characteristic of contemporary health economics and outcomes research utilities, QALYs, cost-effectiveness analysis, mapping, decision modelling and probabilistic analysis are accepted while the logically prior requirements of measurement receive little explicit recognition.

This distinction is fundamental. The interrogation is not asking whether Latin American researchers are technically proficient in economic evaluation. There is extensive evidence of precisely such proficiency. The question is whether the knowledge environment makes lawful measurement a prerequisite for the arithmetic used to support quantitative claims of therapy impact.

On the standards stipulated by the 24-item interrogation, it does not. This is measurement inversion: arithmetic precedes measurement. Where quantitative therapy-impact claims depend upon that inversion, the consequence is measurement failure.

Measurement Precedes Arithmetic

The pivotal true statement, measurement precedes arithmetic, receives a low endorsement ($p = 0.15$). The equally fundamental proposition that meeting the axioms of representational measurement is required before arithmetic receives the same result ($p = 0.15$).

These findings capture the central feature of the Latin American knowledge base. The literature and methodological environment emphasize what analysts should do with numbers: estimate utilities, construct QALYs, calculate incremental cost-effectiveness ratios, undertake sensitivity analyses and populate decision models. Far less attention is given to the prior scientific question: what empirical attribute does these numbers represent and what operations does their scale permit?

Mathematical correctness cannot answer that question. A calculation may be executed perfectly while remaining empirically meaningless if its operands do not possess the required measurement properties.

Unidimensionality and Attributes

The proposition that measures must be unidimensional receives only p = 0.20. This points to a more fundamental problem than incorrect arithmetic because measurement begins with identification of the attribute.

Contemporary HEOR frequently starts instead with an instrument. Multiattribute health-status systems describe mobility, pain, anxiety, self-care and other dimensions and then combine them into a numerical index. Yet combining dimensions does not establish a single measurable attribute.

The false proposition that time trade-off preferences are unidimensional receives strong endorsement at p = 0.85. Eliciting a numerical preference does not establish that a single latent attribute has been identified, much less measured. Measures are for single attributes.

The sequence has again been reversed. Rather than defining an attribute and constructing an appropriate measurement system, a numerical score is produced and subsequently treated as though its existence demonstrates measurement.

Ratio Measurement and Multiplication

The true proposition multiplication requires a ratio measure receives only p = 0.15. This has immediate implications for cost-utility analysis.

The interrogation strongly endorses time as a ratio measure (p = 0.95). Duration has a meaningful zero and supports meaningful ratios. The difficulty arises with the quantity by which time is discounted (or multiplied) to create a QALY.\.

A health-state preference or utility score does not acquire ratio properties simply because analysts wish to use it as a multiplier. The required scale properties must exist before multiplication takes place. The Latin American knowledge base therefore reproduces the characteristic inversion of the international QALY framework: multiplication is institutionally accepted ratio status of both operands has been established.

The QALY

The QALY items provide the most visible expression of this measurement inversion. The knowledge base strongly endorses the false proposition that the QALY is a ratio measure (p = 0.90), the further false proposition that it is dimensionally homogeneous (p = 0.90), and the false proposition that QALYs can be aggregated (p = 0.95). Apart from being false propositions, the knowledge base assumes that the QALY is an impossible quantitative measure.

This is a textbook representation of measurement inversion. The numerical construction is accepted first; its measurement status is assumed afterwards.

If the utility component has not independently been established as a lawful ratio measure, applying it as a discount factor is impermissible. The QALY is an impossible construct.

Preference Algorithms, Summed Scores and Mapping

The false proposition that EQ-5D-3L preference algorithms create interval measures receives p = 0.90. The propositions that summations of subjective instrument responses and Likert scores create ratio measures also receive p = 0.90.

These findings expose a broader assumption: numerical transformation is treated as though it can create measurement. It cannot. Regression can predict numbers. An algorithm can generate numbers. Mapping can produce values resembling those of a target instrument. Statistical weighting can produce an index. None of these procedures establishes the empirical representation required for measurement.

This distinction is particularly important for mapping. A statistically successful relationship between two sets of scores does not transfer the measurement properties of one scale to another. An ordinal or otherwise non-ratio source does not become ratio merely because its values predict an outcome expressed on a different numerical scale. Prediction is not measurement transformation.

Negative Values and the True Zero

The false proposition ratio measures can have negative values receives p = 0.90. This illustrates the disconnect between conventional utility practice and the requirements attributed to ratio measurement in this framework. Preference-based systems can assign negative values to health states interpreted as worse than death. Those values can subsequently enter QALY calculations.

The issue is not whether negative numbers are mathematically permissible. They plainly are. The issue is whether a scale containing such values can simultaneously possess the meaningful non-arbitrary zero required for the ratio interpretation being claimed. Assigning the numeral zero a semantic label does not establish an empirical zero.

Rasch Measurement and Latent Attributes

The strongest negative logits occur among the Rasch-related propositions. Three central propositions receive p = 0.05 and −2.50 logits: the stipulated distinction between linear ratio and Rasch logit ratio measurement, the transformation of subjective responses under Rasch requirements, and the role attributed to Rasch measurement in assessing latent therapy-impact attributes. The relationship between Rasch rules and the stipulated representational-measurement axioms is similarly weakly endorsed.

Possession of the latent attribute receives only p = 0.20.

These are probably the most damaging responses as Rasch measurement, the Rasch logit ratio scale, is the only measure allowed following the axioms of representational measurement for latent attributes. Subjective responses must meet Rasch measurement standards. There is no alternative.

Falsifiability and Reference-Case Simulation

The results are somewhat less negative for falsifiability. The proposition that non-falsifiable claims should be rejected receives p = 0.65.

Yet the false statement that reference-case simulations generate falsifiable claims receives p = 0.85. This reveals another form of conceptual inversion: producing a numerical model result is easily confused with producing an empirically testable scientific claim.

Models are conditional constructions. Their outputs depend upon assumptions, parameters, structural choices and frequently extrapolations over periods for which direct observations do not exist. Moreover, if the model incorporates inputs whose measurement properties have not been established, computational sophistication simply carries those deficiencies forward.

A Regional Pattern not a Latin American Invention

This is perhaps the most important interpretation of the findings.

The interrogation should not be presented as demonstrating a peculiarly Latin American failure. Latin American HEOR has developed within a mature international framework in which cost-effectiveness analysis, utilities, QALYs, preference-based instruments, modelling and sensitivity analysis are already presented as established methodological practice. The fact that this is only possible if representational measurement is ignored guarantees measurement failure.

ISPOR is important precisely because it provides professional networks, educational activities, conferences, publications and methodological guidance through which this international framework can be disseminated. The fact that what is being presented fails the standards for lawful measure is entirely overlooked. This is not the basis for the quantitative assessment of therapy claims.

The relevant question is therefore not: Why did Latin America invent measurement inversion?

It did not. The more important question is: What measurement framework did Latin America inherit? It inherited a framework that teaches false measurement.

Implications for Latin American HTA

The implications are particularly important where national HTA systems are still developing. Methodological choices made during institutional formation can become embedded for decades.

If utilities, QALYs, mapping, ICERs and reference-case modelling are incorporated into guidelines before the requirements of lawful measurement are considered, measurement inversion becomes part of the institutional infrastructure itself. The opportunity is therefore significant. Developing HTA systems need not reproduce every methodological convention inherited from established systems. They can ask the logically prior questions that mature systems failed to ask.

What is the therapy-impact attribute? Is it manifest or latent? What measurement structure is required? What arithmetic is admissible? What empirical claim is being made? How will it be evaluated? What observation would falsify it?

Those questions fundamentally change HTA.

Measurement Inversion as Measurement Failure

Taken together, the 24 items produce a coherent diagnostic signature. Statements representing the stipulated foundations of lawful measurement generally receive weak endorsement. Statements representing conventional HEOR numerical practices receive strong endorsement.

The calculations undertaken within contemporary HEOR may be mathematically correct. That is not the issue. Mathematics establishes whether arithmetic has been performed correctly. Measurement establishes whether performing that arithmetic on those numbers has empirical meaning.

Under the standards applied in this interrogation, the ISPOR-related Latin America knowledge base reverses those priorities. The conclusion therefore parallels Africa and Asia: measurement inversion is measurement failure.

THE TRANSITION TO LAWFUL MEASUREMENT

If the purpose of health technology assessment is to evaluate therapy impact, the measurement requirement comes first. Manifest attributes require lawful linear ratio measures. For latent attributes the solution is Rasch logit ratio measurement. Only after lawful measurement has been established can arithmetic support evaluable and falsifiable claims. Without that foundation, increasing mathematical sophistication does not produce better measurement. It produces increasingly sophisticated numerical storytelling.

The alternative is not another layer of statistical refinement but a return to measurement. Transition must begin by providing faculty and practitioners with the knowledge and practical tools required to reconsider the scientific foundations of what they teach and apply. It is precisely this gap that the Maimon Research nine-unit transformation program is intended to address [5]. Designed specifically for HTA, the program provides a structured transition from the legacy framework represented by the ISPOR Good Practices recommendations to one grounded in representational measurement. Beginning with the principle that measurement precedes arithmetic, the nine units progressively address the definition of attributes, scales of measurement, admissible arithmetic, dimensional homogeneity, manifest and latent attributes, linear ratio measurement, Rasch measurement, attribute possession and the construction of empirically evaluable therapy-impact claims.

The purpose is not to append a measurement module to contemporary HTA while leaving its existing analytical structure intact. It is to demonstrate how the entire discipline changes once representational measurement becomes its organizing scientific principle. Utilities, mapping, QALYs and reference-case simulation are no longer taken as starting points. The starting point becomes the attribute for which a therapy-impact claim is required and the measurement system necessary to support that claim. Arithmetic follows measurement rather than determining retrospectively what is to be treated as a measure.

The nine-unit program then carries measurement into practice. Manifest therapy-impact attributes are addressed through lawful linear ratio measures, while latent attributes are addressed through Rasch measurement with requirements for unidimensionality, invariance, specific objectivity and possession of the defined attribute. These measures provide the foundation for prospectively specified claims and protocols defining the target population, comparator, attribute, measurement instrument, timeframe, success criterion and falsification rule. Therapy impact can then be evaluated against empirical observations rather than inferred primarily from constructed numerical models.

The need for transition is also institutional. Measurement inversion has persisted because the inherited framework is reproduced through teaching, methodological guidance, professional standards, journals, HTA agencies and international networks. Correcting measurement failure therefore requires more than persuading individual analysts to modify particular techniques. The sequence of training itself must change. New practitioners should encounter attributes, scale properties and admissible arithmetic before utilities, QALYs and cost-effectiveness modelling. Established practitioners must similarly be given the opportunity to reconsider familiar methods against measurement standards that should have preceded them. Without this change, each new generation risks reproducing the same measurement failure with increasingly sophisticated analytical tools.

The Program's objective is therefore not simply to demonstrate why the ISPOR Good Practices framework fails to satisfy the requirements of lawful measurement. It provides a coherent replacement framework for teaching, research and HTA practice. The transition is conceptually straightforward: define the attribute; establish the appropriate lawful measure; determine the arithmetic admissible for that measure; formulate the therapy-impact claim; specify the protocol; observe the outcome; and subject the claim to empirical evaluation and potential falsification. After forty years in which arithmetic has been permitted to precede measurement, the task is not to refine the inherited framework. It is to replace measurement inversion with the scientific sequence required for quantitative claims: measurement first, arithmetic second.

CONCLUSION

The interrogation of the ISPOR Latin America Consortium knowledge base identifies a consistent pattern of measurement inversion. Foundational propositions concerning measurement preceding arithmetic, unidimensionality, ratio measurement, admissible arithmetic, representational measurement and the measurement of latent attributes receive weak endorsement. In contrast, propositions supporting the conventional quantitative framework of contemporary HEOR including utilities, QALYs, preference algorithms, aggregation, cost-effectiveness analysis and reference-case modelling receive substantially stronger endorsement. The resulting knowledge environment privileges numerical manipulation without first requiring demonstration that the numbers being manipulated possess the measurement properties necessary to support those operations.

The significance of this finding extends beyond ISPOR itself. Latin America encompasses countries with established HTA institutions as well as countries developing or expanding formal capabilities for health technology assessment, economic evaluation, pricing, reimbursement and healthcare priority setting. The danger is therefore one of institutional reproduction. If developing HTA systems adopt the established international HEOR framework without first examining its measurement foundations, measurement inversion becomes embedded in methodological guidelines, university curricula, professional training and reimbursement processes. Increasing technical sophistication may then reinforce rather than correct the underlying measurement failure.

This should not be interpreted as a Latin American failure. The Latin America Consortium operates within an international methodological environment that has promoted utilities, QALYs, cost-effectiveness analysis, preference-based instruments, mapping and simulation modelling as established methods of quantitative evaluation. The regional findings are therefore better understood as evidence of the successful transmission of an inherited international framework whose measurement foundations were never adequately established. Latin America has inherited the problem; it did not create it.

The required response is not the abandonment of quantitative evaluation but its reconstruction around lawful measurement. Therapy-impact assessment must begin with the attribute to be measured, followed by establishment of the appropriate measurement system and determination of the arithmetic operations that the resulting measure can legitimately support. Manifest attributes require lawful linear ratio measures; latent attributes require invariant Rasch logit ratio measurement. Only after these measurement requirements have been satisfied should arithmetic operations be undertaken and quantitative claims advanced. Claims should then be prospectively specified, empirically evaluated, replicated and exposed to falsification.

For the ISPOR Latin America Consortium, this presents an important choice. The region can continue to reproduce a methodological framework inherited from contemporary international HEOR, thereby embedding measurement inversion more deeply as HTA capacity expands, or it can reconsider the scientific foundations upon which that expansion is based. The opportunity is to place measurement before arithmetic and establish lawful measurement as the organizing principle for future HTA. Doing so would not represent a retreat from quantitative science. It would represent the transition to 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] Wright B. Solving measurement problems with the Rasch Model. *J Educational Measurement.* 1977;14(2):97-116

[5] Maimon Research. HTA Reconstruction. https://maimonresearch.com/hta-reconstruction-program-and-fees/