MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

THE MEASUREMENT FUTURE IN HEALTH TECHNOLGY ASSESSMENT

LEGACY HTA: ISPOR – THE ENDORSEMENT OF MEASUREMENT INVERSION IN PRACTICE PATTERN STANDARDS

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

LOGIT WORKING PAPER No 2000 AUGUST 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

Background: The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) occupies an influential position in defining and disseminating methodological standards for health economics, outcomes research and health technology assessment (HTA). Through its Good Practices reports, task forces, education and professional networks, ISPOR has contributed to the international institutionalization of cost-effectiveness analysis, utilities, quality-adjusted life years (QALYs), preference measurement, patient-reported outcomes and decision modelling. The scientific status of these practices depends, however, upon whether the quantities entering their arithmetic satisfy the requirements of representational measurement.

Objectives: To assess whether the ISPOR Good Practices knowledge base recognizes and reinforces the measurement principles required to determine whether the arithmetic of contemporary HTA is admissible.

Methods: The publicly represented ISPOR methodological knowledge environment was interrogated using 24 canonical propositions derived from representational measurement theory and Rasch measurement principles. Each proposition was classified objectively as TRUE or FALSE. Categorical endorsement probabilities estimated the extent to which the knowledge base explicitly or implicitly reinforced each proposition. Probabilities were transformed to normalized logits within a ±2.50 range.

Results: A pronounced measurement inversion was identified. TRUE propositions fundamental to lawful arithmetic were weakly possessed: measurement precedes arithmetic 0.10 (−2.20 logits); multiplication requires ratio measurement 0.10 (−2.20); and representational-measurement axioms are required for arithmetic 0.10 (−2.20). Conversely, FALSE propositions supporting legacy HTA were strongly reinforced: ratio measures can have negative values 0.90 (+2.20); EQ-5D preference algorithms create interval measures 0.90 (+2.20); the QALY is a ratio measure 0.95 (+2.50); and QALYs can be aggregated 0.95 (+2.50). Rasch propositions required for latent-attribute measurement were predominantly 0.05 (−2.50). The FALSE proposition that reference-case simulations generate falsifiable claims received 0.90 (+2.20).

Conclusions: The ISPOR Good Practices knowledge base strongly possesses the arithmetic of legacy HTA while weakly possessing the measurement requirements necessary to justify it. This is particularly consequential because ISPOR does not merely apply established methods; it helps legitimize, standardize and transmit them internationally. Professional consensus, collective task-force authorship and international repetition cannot establish measurement properties. Good research practice requires the sequence attribute → measurement → scale → admissible arithmetic → quantitative claim → empirical evaluation, replication and falsification. ISPOR Good Practices require reconstruction around a fundamental scientific requirement: measurement must precede arithmetic.

INTRODUCTION

The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) occupies a distinctive position in the development and international transmission of health technology assessment (HTA), health economics and outcomes research methodology. Through its Good Practices for Outcomes Research reports, task forces, educational activities, publications, conferences and associated professional networks, ISPOR has contributed substantially to defining what constitutes accepted methodological practice in economic evaluation, decision modelling, patient-reported outcomes, preference measurement and related fields. Its influence extends well beyond its own membership. ISPOR practice standards and methodological conventions have helped shape the analytical environment within which researchers are trained, manufacturers prepare evidence, consultants construct models, journals evaluate submissions and HTA agencies assess claims for healthcare technologies.

This influence creates a correspondingly important scientific responsibility. A practice standard does more than describe what analysts commonly do. By identifying an approach as good practice, it legitimizes that approach, encourages its reproduction and contributes to its transmission to subsequent generations of researchers. The relevant question is therefore not simply whether ISPOR recommendations are widely accepted, professionally influential or technically sophisticated. It is whether the quantitative operations endorsed within that methodological environment satisfy the measurement requirements necessary to give those operations empirical meaning [i] [ii] .

This question precedes economic evaluation itself. Quantitative science does not begin with arithmetic. It begins with an attribute and the establishment of measurement. Numerical values cannot legitimately be added, subtracted, multiplied or divided merely because numbers have been assigned to observations or generated by an algorithm. The measurement properties of the quantities must first be established and the arithmetic proposed for them shown to be admissible. The distinction between nominal, ordinal, interval and ratio scales is therefore fundamental. Multiplication and division require ratio measurement, with equal units and a non-arbitrary zero, while meaningful quantitative operations must remain invariant under the admissible transformations of the scale concerned.

These requirements have immediate implications for the methodological practices promoted within contemporary HTA. Cost-utility analysis multiplies health-state utility values by time to create quality-adjusted life years (QALYs). Incremental cost-effectiveness analysis subtracts QALYs and subsequently divides differences in costs by differences in QALYs. Multiattribute preference instruments assign numerical values to combinations of responses across different health dimensions. Patient-reported outcome instruments frequently assign numbers to ordered categories, sum responses or statistically transform scores. Decision models subsequently combine costs, probabilities, durations, utilities, survival estimates and other numerical inputs through extensive sequences of arithmetic operations.

The scientific question is logically prior to every one of these operations: have the quantities entering the arithmetic been demonstrated to possess the measurement properties that the arithmetic requires? The QALY provides the clearest example. Its familiar construction, QALY

= U × T requires multiplication. Time can satisfy the requirements for linear ratio measurement. The critical issue is therefore the measurement status of the utility value. If utility is to function as a dimensionless proportional adjustment to time, the measurement properties necessary for that role must be demonstrated independently. They cannot be conferred retrospectively by performing the multiplication. Nor does describing a health-state value as a utility establish that it possesses equal units, invariance and a non-arbitrary zero.

This becomes particularly problematic where preference-based health-state values can take negative values while simultaneously being treated as though they possess ratio properties. A ratio scale requires a non-arbitrary zero representing absence of the measured attribute. Numerical values extending below that zero are incompatible with the ratio interpretation required for meaningful proportional comparisons. Yet the arithmetic of cost-utility analysis proceeds as though the required ratio structure had already been established.

The failure, if present, cannot be repaired downstream. If utility has not been established as an admissible quantity for discounting time, the resulting QALY does not acquire measurement legitimacy merely because a numerical product has been calculated. Subtracting QALYs to create an incremental QALY cannot manufacture the missing measurement properties. Dividing incremental costs by that difference cannot subsequently transform the denominator into a lawful measure. Probabilistic sensitivity analysis, scenario analysis, microsimulation and increasingly sophisticated decision models cannot correct a measurement failure that occurred before the modelling began. Arithmetic cannot rescue a failure of measurement by adding further arithmetic.

The same requirement applies to latent attributes. Pain, fatigue, physical function, treatment satisfaction, need fulfilment and other patient-reported attributes are not directly observable quantities. Assigning integers to ordered response categories and summing them creates a score; it does not establish measurement. A measurement pathway must first demonstrate a defined unidimensional attribute and construct an invariant quantitative continuum upon which possession of that attribute can be located. Rasch measurement provides the relevant person-item conjoint measurement framework, with specific objectivity and invariance providing the basis for quantitative assessment of latent attributes [iii] [iv] . Statistical sophistication applied to ordinal scores cannot substitute for the construction of a measurement system.

There is no distinctive feature of HTA, pharmacoeconomics or outcomes research that provides an exemption from these requirements. It might be argued that economic evaluation is intended as an indicative decision tool rather than as physical measurement. That distinction does not resolve the problem. A decision tool that assigns numbers to attributes, combines those numbers, multiplies them by time, aggregates them across populations and divides costs by estimated outcomes is making quantitative claims. Its purpose does not alter the measurement requirements governing the arithmetic used in its construction. Decision usefulness cannot substitute for measurement validity.

This point is especially important for ISPOR. National HTA agencies may inherit established methodologies and apply them within jurisdiction-specific decision processes. ISPOR occupies a different position. Through its practice standards and associated methodological activities, it participates directly in defining and disseminating the standards by which those methodologies are

judged. If representational measurement is absent from that framework, the consequences are therefore not confined to a single method, task force or jurisdiction. The omission can be reproduced internationally through guidance, education, research, publication and professional practice.

The present assessment revisits the ISPOR Good Research Practice knowledge base from this perspective. The unit of analysis is not the competence, intentions or personal beliefs of individual ISPOR members, task force participants or authors. It is the publicly represented methodological knowledge environment created by the recurring assumptions, recommendations and quantitative conventions embodied across ISPOR practice standards and related activities. The question is whether that knowledge environment recognizes and applies the requirements of representational measurement before endorsing the arithmetic of contemporary health-economic evaluation.

The assessment employs the same 24 canonical measurement propositions used across the wider HTA knowledge base measurement inversion interrogation program. Each proposition has a defined TRUE or FALSE status under the representational-measurement framework. Categorical endorsement probabilities assess the extent to which the ISPOR knowledge environment explicitly or implicitly reinforce each proposition, with probabilities transformed to normalized logits to make the pattern of possession and non-possession transparent. A high endorsement probability for a FALSE proposition does not imply that ISPOR members consciously believe something false. It indicates that the methodological knowledge environment behaves as though the proposition were acceptable through its recommendations, assumptions or repeated quantitative practices.

The distinction is critical because ISPOR's influence makes this more than an assessment of one professional organization's methodological preferences. If a body that has helped define international good practice strongly possesses the arithmetic of utilities, QALYs, cost-effectiveness analysis, preference measurement and decision modelling while failing to possess the measurement requirements necessary to justify that arithmetic, then the problem is one of standards themselves.

Professional consensus cannot establish a measurement scale. Publication cannot establish a non-arbitrary zero. A task force cannot confer ratio properties upon an ordinal or otherwise unqualified numerical representation. Repetition across guidelines cannot make multiplication admissible. International adoption cannot transform convention into measurement.

The issue addressed in this paper is therefore straightforward but fundamental: does ISPOR Good Research Practice begin where quantitative science must begin with measurement or has it helped institutionalize a global methodological framework in which arithmetic was allowed to precede measurement?

If the latter is demonstrated, then the appropriate response is not another refinement of legacy HTA methodology. It is reconsideration of what the term good research practice should require.

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 GOOD PRACTICE KNOWLEDGE BASE

For the purposes of this assessment, the ISPOR Good Practices knowledge base is defined as the cumulative methodological environment created through ISPOR task-force reports, Good Practices for Outcomes Research recommendations, educational activities, conference outputs and related methodological guidance. It is not represented by any single publication, committee or group of authors. Rather, it is the recurring body of concepts, assumptions, quantitative conventions and recommended practices through which ISPOR has contributed to defining acceptable research practice in pharmacoeconomics, outcomes research and health technology assessment.

This distinction is important. Individual ISPOR reports typically address particular methodological questions and are produced by different groups of experts. The relevant knowledge base emerges from their cumulative effect. Where the same concepts and quantitative practices recur across economic evaluation, decision modelling, preference measurement, patient-reported outcomes and other methodological areas, they constitute a recognizable professional framework within which analysts are encouraged to design studies, construct models, interpret outcomes and communicate quantitative claims.

The reach of this knowledge base is substantial because ISPOR operates across several interconnected methodological domains. These include cost-effectiveness and cost-utility analysis, decision-analytic modelling, economic evaluation, preference-based health-state assessment, patient-reported outcomes, value assessment and the treatment of uncertainty. The resulting knowledge environment therefore encompasses much of the quantitative architecture of contemporary HTA rather than a narrowly defined area of research practice.

A central feature of this environment is its emphasis upon methodological standardization. Good-practice recommendations seek to improve consistency, transparency and technical quality by specifying how analyses should be designed, conducted, documented and reported. In modelling, attention is directed toward model structure, assumptions, parameterization, validation, uncertainty and sensitivity analysis. In economic evaluation, costs and outcomes are combined to produce comparative estimates intended to support decision-making. In outcomes research, numerical representations of health status and patient experience provide inputs to clinical, economic and value assessments.

The knowledge base is consequently highly arithmetic. Probabilities, costs, survival, durations, health-state values and other numerical inputs are combined within models. Health-state preference values are treated quantitatively and combined with time to generate QALYs. QALYs can then be accumulated across modelled health states and time periods, differences between alternatives calculated and incremental outcomes related to incremental costs through cost-effectiveness ratios. These operations form an established analytical language through which treatment value is represented.

Preference-based multiattribute instruments occupy an important position within this structure. Responses describing health across multiple domains are used to define health states, preferences for those states are elicited or estimated and algorithms assign numerical health-state values. These values can subsequently enter economic models as utilities. Once incorporated into cost-utility analysis, they are treated as quantities capable of multiplication by time and aggregation into QALYs.

Patient-reported outcomes constitute another major component of the knowledge environment. Responses to ordered categories may be numerically coded, combined into scores, summarized and subjected to statistical analysis. Questions of instrument development, reliability, validity, responsiveness and interpretation are therefore important components of outcomes-research practice. The resulting scores may be used to characterize treatment effects, patient experience and changes in health status.

The knowledge base also places substantial emphasis upon decision-analytic modelling. Models provide a framework within which evidence from different sources can be combined, outcomes extrapolated beyond observed periods and alternative assumptions explored. Deterministic sensitivity analysis, probabilistic sensitivity analysis and scenario analysis provide mechanisms for examining uncertainty surrounding model inputs and assumptions. Reference-case conventions can further standardize analytical choices and facilitate comparisons across evaluations.

Taken together, these practices constitute a coherent and mature methodological system. They specify how evidence should be assembled, how numerical inputs should be incorporated, how uncertainty should be represented and how outputs should be reported. The system therefore provides analysts with extensive guidance on what should happen after numerical values have entered the analytical framework.

The present interrogation asks a logically prior question: what must be established before those numbers are permitted to enter the arithmetic?

This is where representational measurement becomes relevant. The existence of a numerical value does not establish its measurement status. A preference algorithm producing a number does not establish an interval or ratio scale. Summation of ordered responses does not demonstrate equal units. Anchoring a health-state value at zero does not establish a non-arbitrary zero. Multiplication of a health-state value by time does not demonstrate that the health-state value possessed the ratio properties necessary for multiplication before the operation was performed.

The distinction between measurement and numerical representation therefore provides the boundary condition for assessing the ISPOR knowledge base. Good practice in model construction, uncertainty analysis, validation and reporting cannot determine whether the quantities entering those procedures are measures unless the required measurement properties have first been established.

For manifest attributes such as survival time, treatment duration, hospital days or walking distance, the relevant question is whether an appropriate linear ratio measure with a defined unit and non-arbitrary zero has been established. For latent attributes such as pain, fatigue, physical function or

need fulfilment, a different measurement pathway is required. Numerical responses must be shown to represent possession of a defined unidimensional attribute and to satisfy the requirements necessary for construction of an invariant quantitative continuum. Within the framework applied in this assessment, Rasch measurement provides that pathway.

The ISPOR Good Practices knowledge base is therefore interrogated not for whether it uses numbers extensively, it clearly doe, but for whether it distinguishes sufficiently between numbers, scores and measures before prescribing arithmetic operations upon them. This distinction is fundamental because methodological guidance can be technically sophisticated at every subsequent stage while remaining scientifically compromised if the measurement status of its inputs was never established.

The same issue applies to empirical claims. Decision models can be internally validated, assumptions varied and uncertainty extensively characterized without establishing that their principal outcome measures possess the properties required for the arithmetic performed upon them. Sensitivity analysis evaluates the consequences of changing model assumptions and inputs; it does not itself establish measurement and does not convert a conditional simulation into an independently falsifiable empirical claim.

The ISPOR knowledge base is particularly important because its methodological influence is transmissive. Good-practice recommendations do not remain confined to the documents in which they appear. They become part of the wider professional environment through education, research, consulting, publication, manufacturer submissions, academic training and interaction with HTA agencies. Repeated methodological conventions can therefore become increasingly embedded across institutions and jurisdictions.

For this reason, the unit of analysis in the present assessment is explicitly **the knowledge environment rather than the individual**. No inference is made about what individual ISPOR members, task-force authors or researchers personally understand or believe about measurement theory. An endorsement probability represents the extent to which the publicly represented knowledge base behaves as though a proposition is accepted, implied or reinforced through its recurring methodological practices.

This distinction also prevents professional authority from being confused with scientific demonstration. The participation of experienced researchers, economists, statisticians, psychometricians and decision scientists can contribute substantially to the technical sophistication of practice guidance. It cannot, by itself, establish the measurement properties of the quantities employed. Those properties must be demonstrated independently.

The ISPOR Good Practices knowledge base can therefore be characterized as an extensive, internationally influential and highly developed system for the construction and application of quantitative health-economic and outcomes analyses. The scientific issue addressed by the present interrogation is narrower but more fundamental: does that system possess and apply the measurement principles required to determine whether its established arithmetic is admissible?

That question must be answered before the designation good practice can itself be evaluated.

.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. Each statement is objectively TRUE or FALSE under standards 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 OF THE ISPOR GOOD PRACTICES KNOWLEDGE BASE

The interrogation of the ISPOR Good Practices knowledge base presents a remarkably coherent picture of measurement inversion (Table 1). The significance of the findings lies not simply in

the endorsement or rejection of individual propositions, but in the direction of the profile as a whole. Statements that express the measurement requirements necessary to constrain arithmetic are weakly possessed, frequently with strongly negative normalized logits, while false statements that permit the established arithmetic of utilities, QALYs, aggregation and cost-effectiveness analysis are strongly reinforced. The result is a knowledge environment in which the arithmetic of contemporary HTA is substantially better established than the measurement principles required to determine whether that arithmetic has empirical meaning.

TABLE 1: ITEM STATEMENT, RESPONSE, ENDORSEMENT AND NORMALIZED LOGITS ISPOR GOOD RESEARCH PRACTICE

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.15	-1.75
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.75
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
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	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
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	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.70	+0.85
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
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 BE COMBINED WITH A LOGIT SCALE	0	0.65	+0.60
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.15	-1.75
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The starting point is the most fundamental requirement of quantitative science: measurement must precede arithmetic. This TRUE proposition receives an endorsement probability of only 0.10, equivalent to −2.20 normalized logits. The equally fundamental TRUE proposition that meeting the axioms of representational measurement is required before arithmetic is undertaken receives the same 0.10 (−2.20), while the TRUE proposition that multiplication requires ratio measurement is again only 0.10 (−2.20). Recognition that measures must be unidimensional is similarly weak at 0.15 (−1.75), as is recognition that interval measures lack a true zero, also 0.15 (−1.75). Taken together, these findings indicate that the basic principles required to determine whether numerical operations are admissible do not function as effective constraints within the ISPOR Good Practices knowledge environment.

This is a serious finding because ISPOR guidance is not concerned merely with descriptive statistics. It addresses analytical methods that depend upon extensive arithmetic. Utilities are multiplied by time, QALYs are accumulated and subtracted, incremental outcomes are related to

incremental costs, multiattribute health-state values are generated algorithmically, patient-reported responses are scored and aggregated, and decision models combine numerous numerical inputs. If the knowledge system strongly possessed the arithmetic while only weakly possessing the principles governing whether that arithmetic is admissible, then methodological sophistication has developed downstream of an unresolved measurement problem.

The clearest evidence is provided by ratio measurement. The TRUE proposition that multiplication requires a ratio measure receives only 0.10 (−2.20), while the FALSE proposition that ratio measures can have negative values receives 0.90 (+2.20). These two results should be considered together. The first effectively removes the scale restriction that should govern multiplication; the second abandons the non-arbitrary zero required for ratio measurement. The combination creates precisely the permissive measurement environment required for conventional QALY arithmetic. Numerical health-state values may extend below zero while simultaneously being treated as though they can operate proportionally when multiplied by time.

The problem is not that ISPOR guidance explicitly announces that the axioms of ratio measurement should be rejected. The problem is more fundamental: its methodological practices behave as though those axioms need not constrain the arithmetic. A numerical value can enter a calculation because it is conventionally designated a utility, rather than because its measurement properties have first been demonstrated. The distinction between numerical assignment and measurement has therefore effectively disappeared at the point where it matters most.

This becomes unmistakable in the treatment of the QALY. The FALSE proposition that the QALY is a ratio measure receives an endorsement probability of 0.95 (+2.50), the maximum positive value in the normalized interrogation. At the same time, the TRUE proposition that time is a ratio measure receives 0.95 (+2.50). Recognition of the ratio status of time is entirely appropriate, but it exposes rather than resolves the QALY problem. The construction QALY = U × T requires both recognition of the ratio status of time and demonstration that the utility value possesses the measurement properties necessary to function as a dimensionless proportional adjustment to time. Yet the requirement that multiplication demands ratio measurement receives only 0.10 (−2.20). The knowledge base therefore strongly recognizes the ratio status of one component, weakly recognizes the scale requirement governing their multiplication, and nevertheless maximally reinforces the false conclusion that the resulting QALY is itself a ratio measure.

This is measurement inversion in an unusually concentrated form. The arithmetic is accepted first and the measurement status of its product appears to be presumed from the established use of the operation. But multiplication cannot confer measurement properties upon its operands or upon the product it generates. The fact that a numerical calculation can be performed does not establish that it represents an admissible empirical relationship. The measurement properties must exist before the operation is undertaken.

The treatment of preference-based health-state values reinforces the same conclusion. The FALSE proposition that time trade-off preferences are unidimensional receives 0.85 (+1.75), while the FALSE proposition that EQ-5D-3L preference algorithms create interval measures receives 0.90 (+2.20). These endorsements indicate a knowledge environment in which elicitation procedures and numerical algorithms are treated as though they can create measurement properties. They

cannot. A time trade-off procedure can elicit preferences, but preference elicitation does not demonstrate possession of a unidimensional measurable attribute. An EQ-5D valuation algorithm can assign numerical values to health states, but assigning numbers does not establish equal intervals. Anchoring values at selected points does not establish a non-arbitrary zero, and describing the resulting values as utilities does not establish ratio status.

The consequence is that the utility enters cost-utility analysis carrying measurement properties that have been assumed rather than demonstrated. Once admitted, the value is multiplied by time, accumulated across health states, averaged across groups and subjected to further arithmetic. The resulting calculations may be technically impeccable conditional upon the inputs, but technical correctness of calculation cannot compensate for failure to establish that the inputs support the calculations in the first place.

The same conflation between numerical representation and measurement appears in the treatment of subjective responses. The FALSE proposition that summations of subjective instrument responses are ratio measures receives 0.85 (+1.75), while the FALSE proposition that summation of Likert question scores creates a ratio measure receives 0.90 (+2.20). These are strong positive endorsements of propositions that measurement theory rejects. Assigning integers to ordered categories and summing them creates a score. It does not establish equal units, invariance, additivity in the measurement-theoretic sense or a non-arbitrary zero. Reliability, responsiveness, internal consistency, statistical significance and predictive validity may be useful properties of an instrument or score, but none demonstrates that the resulting numerical representation has become a measure capable of supporting unrestricted arithmetic.

The distinction between a number, a score and a measure should therefore be fundamental to any statement of good research practice. The ISPOR profile suggests that it is not. Numerical representation has been allowed to substitute for measurement demonstration, after which the resulting scores and indices become available for progressively more elaborate quantitative manipulation.

The QALY results show what happens downstream. The FALSE proposition that the QALY is dimensionally homogeneous receives 0.85 (+1.75), while the FALSE proposition that QALYs can be aggregated receives 0.95 (+2.50). Both operations presume measurement properties that have not been established. If a utility value has not been demonstrated to possess the structure required to operate as a dimensionless proportional adjustment to time, dimensional homogeneity cannot simply be assumed because the resulting quantity has been named a QALY. Similarly, aggregation presupposes that the quantities being added possess legitimate and commensurable quantitative units. Addition cannot create those units. If the individual QALY is not established as an admissible measure, summing QALYs across health states, years or populations simply propagates the original measurement failure.

The implications for cost-effectiveness are unavoidable. The TRUE proposition that claims for cost-effectiveness fail the axioms of representational measurement receives only 0.15 (−1.75). Yet the denominator of the conventional incremental cost-effectiveness ratio is generated through precisely the sequence of operations already brought into question. Utility is multiplied by time to create the QALY; QALYs are aggregated; differences in QALYs are calculated; and incremental

costs are divided by the resulting incremental QALYs. If the initial multiplication has not been shown to be admissible, subtraction does not repair the failure and division cannot rescue it. A legitimate monetary numerator cannot confer measurement legitimacy upon an inadmissible denominator. The expression “cost per QALY” may be numerically computable and institutionally familiar while lacking the measurement foundation required for interpretation as a scientifically meaningful ratio.

The Rasch propositions reveal the complementary weakness in ISPOR's treatment of latent attributes. The TRUE proposition that the relevant measurement pathways are linear ratio measurement for manifest attributes and Rasch logit ratio measurement for latent attributes receives only 0.05 (−2.50). The TRUE proposition that transformation of subjective responses to interval measurement requires Rasch measurement rules receives 0.05 (−2.50). Recognition of the Rasch logit ratio scale as the basis for assessing therapy impact for latent traits is again only 0.05 (−2.50), while the proposition linking Rasch measurement requirements to the axioms of representational measurement receives the same 0.05 (−2.50). Recognition that the outcome of interest for a latent trait is possession of that attribute is only 0.15 (−1.75).

The concentration of these propositions at or near the negative extreme is important. ISPOR operates in a field where latent attributes are ubiquitous. Pain, fatigue, physical function, symptoms, satisfaction, quality-of-life domains and need fulfilment cannot be directly observed in the same manner as survival time or distance walked. Responses to questions are observations from which a measurement system must be constructed. If the objective is to make quantitative therapy-impact claims, the relevant requirement is not merely to generate a statistically useful score but to establish a defined unidimensional attribute, invariant measurement structure and a quantitative continuum upon which possession of the attribute can be located. The interrogation indicates that this measurement pathway is essentially absent from the Good Practices knowledge environment while conventional scoring and numerical transformation are strongly embedded.

The contrast is difficult to reconcile with the designation “good practice.” A professional standard concerned with patient-reported outcomes should distinguish explicitly between collecting responses, generating scores and constructing measures. It should require the measurement properties of the outcome to be established before arithmetic claims of treatment effect are made. The ISPOR profile instead indicates that statistical procedures have been allowed to occupy the place that measurement should have occupied.

The results for falsifiability reveal a related inversion. The TRUE proposition that non-falsifiable claims should be rejected receives a comparatively positive endorsement of 0.70 (+0.85). This indicates some recognition of falsifiability as a general scientific requirement. Yet the FALSE proposition that reference-case simulations generate falsifiable claims receives 0.90 (+2.20). The juxtaposition is revealing. Falsifiability appears to be recognized abstractly while its implications for the central modelling practices of HTA are not.

A reference-case simulation is conditional upon model structure, assumptions, parameter estimates, extrapolations and selected input values. Deterministic sensitivity analysis can show how outputs change when selected inputs are varied. Probabilistic sensitivity analysis can propagate assumed parameter uncertainty through the model. Scenario analysis can compare

alternative sets of assumptions. These exercises may be useful for understanding the internal behavior of the model, but they are not equivalent to empirical falsification. Sensitivity analysis asks what happens to an output when assumptions are changed. Falsification asks whether an empirical claim survives confrontation with independently observed evidence. Internal model robustness cannot substitute for empirical risk.

This distinction matters because the sophistication of modelling can easily create an appearance of scientific strength. More parameters, more simulations, more extensive sensitivity analyses and more elaborate validation exercises can make a model technically impressive. None can manufacture measurement properties absent from the inputs, and none can transform a conditional projection into a falsifiable therapy-impact claim simply by increasing computational complexity. A measurement failure that occurs before the model is constructed remains a measurement failure after ten thousand probabilistic simulations.

The overall profile is therefore not one of random methodological gaps. Its direction is too consistent. The TRUE propositions most capable of constraining legacy HTA are weakly represented: measurement preceding arithmetic 0.10 (−2.20); multiplication requiring ratio measurement 0.10 (−2.20); representational-measurement axioms required for arithmetic 0.10 (−2.20); unidimensionality 0.15 (−1.75); the failure of cost-effectiveness claims under those axioms 0.15 (−1.75); and the principal Rasch measurement requirements predominantly 0.05 (−2.50). In the opposite direction, the FALSE propositions necessary to sustain the established framework are strongly reinforced: ratio measures permitting negative values 0.90 (+2.20); EQ-5D preference algorithms creating interval measures 0.90 (+2.20); the QALY as a ratio measure 0.95 (+2.50); Likert summation producing ratio measurement 0.90 (+2.20); QALYs being aggregable 0.95 (+2.50); and reference-case simulations generating falsifiable claims 0.90 (+2.20).

One anomalous item does not alter this pattern. The FALSE proposition that a linear ratio scale for manifest claims can always be combined with a logit scale receives 0.65 (+0.60). It is appropriately treated as an outlier rather than as evidence against the overall interpretation. The central finding does not depend upon every item moving in the same direction. It rests upon the extraordinary consistency of the propositions directly governing measurement, QALY construction, aggregation, cost-effectiveness, latent measurement and falsifiability.

The importance of the ISPOR result is magnified by the nature of the organization. ISPOR is not simply another user of an inherited national HTA methodology. Through its task forces and Good Practice reports it participates in defining, codifying and transmitting professional methodological standards. These reports are commonly produced through collective authorship involving groups of experienced researchers and subject-matter experts. It would be inappropriate to infer from the present interrogation that every task-force member consciously understood or endorsed the measurement inversion identified here. The analysis does not interrogate individual beliefs or intentions. Its unit of analysis is the published knowledge environment.

The collective authorship nevertheless matters. The final reports represent collective professional products carrying ISPOR authority. Measurement inversion has therefore repeatedly survived expert selection, task-force discussion, drafting, review and publication without the measurement requirements governing the recommended arithmetic becoming effective constraints upon that

guidance. The issue is not personal complicity. It is that expert consensus repeatedly failed to function as a measurement safeguard.

This makes the institutional failure more significant rather than less. If the problem could be attributed to one analyst, one paper or one poorly designed model, it would be readily corrected. Instead, the interrogation indicates that the problem has become embedded within a professional knowledge system capable of reproducing itself. Good Practice reports legitimize methods; those methods enter teaching, publications, consultancy, manufacturer submissions and HTA assessments; their repeated use reinforces their professional acceptability; and professional acceptability is subsequently taken as evidence that the methodological framework is established.

But establishment is not validation.

This distinction is particularly important when ISPOR's international influence is considered. Similar utilities, QALYs, cost-effectiveness methods, preference instruments and modelling conventions recur across national HTA systems. Such convergence can create a powerful appearance of independent confirmation. Yet international repetition of a methodological convention is not replication of a scientific result. If national systems have inherited the same methodological assumptions from a common professional environment, their agreement is evidence of successful transmission, not independent measurement validation.

Professional consensus cannot establish a measurement scale. Agreement among task forces cannot establish a non-arbitrary zero. Publication cannot confer ratio properties. A guideline cannot make inadmissible arithmetic admissible. Thousands of cost-effectiveness models cannot supply retrospectively the measurement proof that should have preceded the first multiplication of a health-state value by time.

What has been replicated is the methodology, not the measurement proof.

There is also no unique feature of HTA that provides a scientific exemption from these requirements. Economic evaluation may be defended as an indicative decision tool rather than a system intended to reproduce physical measurement. That distinction does not resolve the problem. An indicative decision tool that assigns numbers to attributes, combines those numbers, multiplies them by time, aggregates them across persons and ultimately divides costs by estimated outcomes is making quantitative claims. The purpose for which the resulting numbers are used does not change the measurement requirements governing the arithmetic used to create them. Decision usefulness cannot substitute for measurement validity.

This is particularly evident in the construction of multiattribute health-state values. Responses across distinct dimensions are numerically coded; combinations of those responses define health states; preferences for those states are elicited; algorithms assign numerical values; those values are designated utilities; utilities are multiplied by time; QALYs are accumulated and compared; and incremental costs are divided by incremental QALYs. At no stage does the succession of numerical operations itself demonstrate a unidimensional attribute, equal units, invariance, a non-arbitrary zero or the ratio properties required for multiplication. Each stage may generate another number. Producing another number is not the same as constructing a measure.

The failure identified in the ISPOR profile therefore occurs before economic modelling begins. Once this is recognized, many of the refinements traditionally associated with good research practice become secondary. Transparency is desirable, but transparent inadmissible arithmetic remains inadmissible. Sensitivity analysis is useful, but uncertainty analysis around an unestablished measure cannot establish the measure. Model validation may demonstrate that a model operates as designed, but it cannot demonstrate that its inputs possess measurement properties they never acquired. Consensus can improve consistency, but consistent repetition of the same measurement failure does not convert failure into validity.

The ISPOR Good Practices profile indicates a knowledge system that has substantially institutionalized a different sequence: numerical representation is accepted, arithmetic follows, models are constructed, uncertainty is explored and decision claims are generated, while the prior questions of measurement and admissible arithmetic remain weakly represented. That is the defining feature of measurement inversion.

The conclusion is consequently more serious than saying that ISPOR Good Practices should include greater discussion of measurement theory. Adding a paragraph on scales of measurement to an otherwise unchanged framework would not address the problem. Measurement must become a gatekeeper. A proposed quantitative operation should not proceed until the attribute has been identified, its measurement structure demonstrated, its scale properties established and the arithmetic shown to be admissible. The good practices need to be audited to assess compatibility with lawful measurement; one example is the good practices for mapping [v] .

ISPOR has helped establish an internationally influential conception of good research practice that endorses measurement inversion. The interrogation indicates that this conception has been extraordinarily successful in standardizing how the arithmetic of legacy HTA should be performed while failing to establish whether any of that arithmetic should have been performed in the first place [vi] [vii]

The distinction is decisive. Good arithmetic performed on quantities that have not been demonstrated to be measures is not good quantitative science. Until measurement precedes arithmetic, the designation *Good Practices* describes professional convention rather than demonstrated measurement validity. Unless they can demonstrate a commitment to lawful measurement they should be abandoned.

CONCLUSION

The interrogation of the ISPOR Good Practices knowledge base identifies a systematic inversion of the requirements of quantitative science. The central propositions that should govern arithmetic are weakly possessed: measurement preceding arithmetic receives an endorsement probability of 0.10 (−2.20 logits), multiplication requiring ratio measurement 0.10 (−2.20), and the requirement to satisfy the axioms of representational measurement before arithmetic 0.10 (−2.20). In contrast, propositions necessary to sustain legacy HTA are strongly reinforced despite being false: ratio measures permitting negative values 0.90 (+2.20), the QALY as a ratio measure 0.95 (+2.50), QALYs being aggregable 0.95 (+2.50), and reference-case simulations generating falsifiable

claims 0.90 (+2.20). The pattern is not an isolated methodological anomaly. It is the signature of measurement inversion.

This finding is particularly consequential for ISPOR because Good Practices do not merely describe established methods. They legitimize, standardize and transmit them. Collective task-force authorship, expert review, publication and international adoption can strengthen professional authority, but none can establish measurement properties. The issue is not whether individual authors knowingly endorsed measurement failure. It is that repeated expert processes failed to make measurement an effective safeguard before arithmetic was designated good practice.

Nor can HTA claim a special exemption because its purpose is decision support. An indicative decision tool that multiplies utilities by time, aggregates QALYs and divides costs by incremental outcomes must still establish that the quantities involved possess the measurement properties required for those operations. Decision usefulness cannot substitute for measurement validity.

The required replacement is neither novel nor complicated. Quantitative claims should follow the sequence expected of measurement science: attribute → measurement → scale → admissible arithmetic → quantitative claim → empirical evaluation, replication and falsification. Manifest attributes require appropriate linear ratio measurement; latent attributes require construction of an invariant measurement continuum through Rasch measurement. Arithmetic begins only after measurement has been established.

ISPOR therefore faces a fundamental choice. It can continue refining an inherited framework in which increasingly sophisticated arithmetic is performed upon quantities whose required measurement properties have not been demonstrated, or it can redefine Good Practices around the principle that should have governed quantitative HTA from the outset.

International repetition cannot resolve that choice. What has been replicated is the methodology, not the measurement proof.

Good research practice must begin before the arithmetic begins.

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

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

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

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

[iv] Bond T, Zi Yan, Heene M. Applying the Rasch Modekl: Fundamental measurement in the human sciences (4th Ed). New York" Routledge, 2021

[v] Langley P. ISPOR: Normal science and the absence of representational measurement in Good Practice – The mapping debacle. Logit Working Paper No 3011. https://maimonresearch.com/logit-working-paper-no-3011-july-2026/

[vi] 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/

[vii] 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/