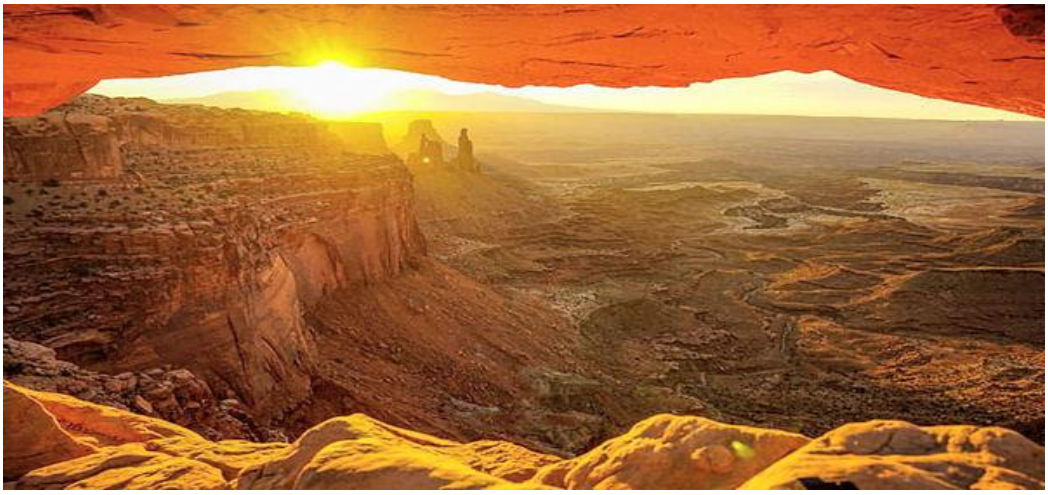
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: THE INTERNATIONAL SOCIETY FOR PHARMACOECONOMICS AND OUTCOMES RESEARCH AS A GUARDIAN OF THE MEMEPLEX OF FALSE MEASUREMENT

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

LOGIT WORKING PAPER No 8 JANUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

This Logit Working Paper series documents a finding as extraordinary as it is uncomfortable: health technology assessment (HTA), across nations, agencies, journals, and decades, has developed as a global system of non-measurement. It speaks the language of numbers, models, utilities, QALYs, "value for money," thresholds, discounting, incremental ratios, extrapolations, and simulations. It demands arithmetic at every turn, multiplication, division, summation, aggregation, discounting, yet it never once established that the quantities to which these operations are applied are measurable. HTA has built a vast evaluative machinery on foundations that do not exist. The probabilities and normalized logits in the country reports that follow provide the empirical confirmation of this claim. They show, with unsettling consistency, that the global HTA.

The objective of this study is to determine whether the International Society for Pharmacoeconomics and Outcomes Research (ISPOR), as the dominant global authority shaping health technology assessment (HTA) standards, guidance, education, and publication norms, operates within a belief system that is compatible with the axioms of representational measurement and the requirements of normal science. Rather than evaluating isolated methodological practices or individual guidelines, the analysis interrogates the underlying epistemic structure of ISPOR's HTA framework. Using a 24-item diagnostic grounded in fundamental measurement theory, the study seeks to identify which propositions are reinforced, tolerated, or excluded within ISPOR's knowledge base, and whether those commitments permit falsification, replication, and the evolution of objective knowledge regarding therapy impact. The aim is not reformist critique but diagnostic clarity: to establish whether ISPOR's quantitative claims are, in principle, capable of being scientific.

The findings are unambiguous and extreme. ISPOR's HTA knowledge base exhibits a near-complete inversion of the scientific order that governs quantitative reasoning. Core axioms of representational measurement—unidimensionality, the requirement of ratio scales for multiplication, the precedence of measurement over arithmetic, and the exclusivity of Rasch measurement for latent traits—are either weakly endorsed or rejected outright, clustering at the floor of the normalized logit scale (−2.50). At the same time, mathematically impossible propositions embedded in conventional HTA practice, QALYs as ratio measures, aggregation of QALYs, summation of ordinal scores as ratio quantities, negative "ratio" utilities, and the falsifiability of reference-case simulations ,are reinforced at or near the ceiling of endorsement (+2.50). This pattern does not reflect confusion, heterogeneity, or debate. It reflects a coherent and stable memeplex in which arithmetic is treated as authoritative while measurement is systematically excluded as a governing constraint. In consequence, ISPOR functions not as a forum for the critical evolution of HTA science, but as a global institutional amplifier of false measurement, conferring professional legitimacy on claims that cannot, even in principle, support falsification or the accumulation of objective knowledge.

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 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 KNOWLEDGE BASE

For the purposes of this analysis, the ISPOR knowledge base is defined as the shared and institutionalized body of concepts, assumptions, evaluative norms, and methodological commitments that are produced, reinforced, and disseminated through ISPOR-sponsored guidelines, task force reports, short courses, conference programming, journal publications, educational materials, and professional accreditation activities. It is not defined by any single document or methodological statement, but by the persistent and recurring patterns that characterize what ISPOR treats as legitimate quantitative reasoning in health technology assessment.

This knowledge base is global in scope and unusually influential. ISPOR functions simultaneously as a professional society, a standard-setting authority, an educational provider, and a publishing hub. Through its Good Practices task forces, value frameworks, conference tracks, short courses, and flagship journals, ISPOR establishes the default vocabulary and analytic boundaries within which HTA is taught, practiced, reviewed, and defended. What ISPOR endorses becomes methodologically "normal"; what it excludes becomes professionally invisible.

The ISPOR knowledge base is anchored in a small set of core methodological commitments that recur across its outputs. These include the routine use of cost-utility analysis, the acceptance of QALYs as valid quantitative outcomes, the treatment of preference-based utility instruments as interval or ratio measures, the legitimacy of aggregating QALYs across individuals and populations, and the use of reference-case simulation models as evidence-generating tools. Sensitivity analysis is treated as a sufficient response to uncertainty, and long-horizon modeled projections are routinely presented as decision-relevant outputs despite the absence of empirical falsifiability.

Equally important are the systematic absences that define the boundaries of this knowledge base. Representational measurement theory is almost entirely absent from ISPOR curricula, task force reports, and methodological guidance. Scale-type constraints defined as nominal, ordinal, interval, ratio and their implications for admissible arithmetic are not treated as gating conditions for analysis. Rasch measurement, despite being the only framework capable of producing invariant measurement for latent traits, is effectively excluded from ISPOR's treatment of patient-reported outcomes. Subjective responses are instead scored, summed, weighted, and transformed through convention rather than measurement.

The ISPOR knowledge base is therefore defined behaviorally rather than philosophically. It is revealed not by explicit statements about measurement, but by what ISPOR repeatedly allows, promotes, and rewards. Composite constructs are treated as measures. Ordinal data are treated as quantitative. Arithmetic is applied without prior demonstration that the quantities involved satisfy the axioms required to support that arithmetic. Simulation outputs are treated as evidence rather

than as conditional projections. Falsifiability is invoked rhetorically but displaced operationally by robustness checks and scenario analysis.

In this sense, ISPOR's knowledge base functions as a closed epistemic system. Its internal coherence is maintained not through empirical testing or theoretical refinement, but through consensus, repetition, and professional reinforcement. Methodological critique that would challenge foundational constructs with utilities, QALYs, ICERs, reference-case models is marginalized, while incremental refinement within the same conceptual architecture is rewarded. The result is a stable belief system that reproduces itself through education, publication, and professional accreditation.

The 24-item diagnostic is therefore applied to ISPOR not as a survey of individual opinions, but as a probe of the conceptual constraints that govern what ISPOR treats as acceptable knowledge. The resulting profile captures the epistemic architecture of ISPOR's HTA framework as it actually operates. As the findings demonstrate, that architecture is incompatible with the axioms of representational measurement and with the requirements of falsification and cumulative scientific knowledge.

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: ISPOR

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.

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

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	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.20	-1.40
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.75	+1.10
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	0.20	-1.40

THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

ISPOR: THE GLOBAL MEMEPLEX OF ARITHMETIC WITHOUT MEASUREMENT

The purpose of this assessment is not to ask whether the International Society for Pharmacoeconomics and Outcomes Research (International Society for Pharmacoeconomics and Outcomes Research) occasionally tolerates weak methods, nor whether some ISPOR members privately harbor doubts about utilities, QALYs, or reference-case modeling. The purpose is to determine whether ISPOR, as the global coordinating institution of HTA, endorses a belief system compatible with scientific measurement. The 24-item diagnostic answers this question decisively. ISPOR is not a neutral convener of debate. It is the primary international stabilizer of a memeplex that inverts the logic of science, placing arithmetic before measurement and treating that inversion as methodological maturity.

This finding is not subtle. It is extreme. Across the diagnostic, propositions that would impose hard constraints on quantitative claims, unidimensionality, scale-type requirements, the precedence of measurement over arithmetic, and the exclusivity of Rasch measurement for latent traits, collapse to the floor of endorsement. At the same time, propositions that are mathematically impossible under representational measurement theory, QALYs as ratio measures, aggregation of QALYs, summation of ordinal scores as ratio quantities, negative "ratio" utilities, cluster at or near the ceiling of endorsement. The resulting logit profile is not mixed or ambiguous. It is polarized. ISPOR occupies the most stable possible equilibrium of arithmetic without measurement.

The first and most consequential result concerns scientific order. The proposition that measurement must precede arithmetic is endorsed at $p = 0.10$, corresponding to a normalized logit of -2.20. This is not a marginal failure. It represents categorical rejection of the most elementary rule of quantitative science. In any discipline that claims to measure, arithmetic is licensed only after the attribute has been shown to possess the relevant structure. ISPOR rejects this ordering outright. Yet it simultaneously endorses arithmetic outputs that presuppose precisely the measurement properties it denies. Aggregation of QALYs is endorsed at $p = 0.95$ (logit $+2.50$). Ratio status of QALYs is endorsed at $p = 0.90$ (logit $+2.20$). Interval status of EQ-5D algorithms is endorsed at $p = 0.85$ (logit $+1.75$). These commitments cannot coexist coherently. ISPOR resolves the contradiction by discarding measurement, not arithmetic.

This inversion explains why ISPOR's reference-case architecture appears so resilient. The incremental cost-effectiveness ratio, lifetime modeling, threshold-based value judgments, and population-level aggregation all depend on ratio-scale arithmetic. Yet the proposition that multiplication requires a ratio measure is endorsed at $p = 0.10$ (logit -2.20). ISPOR therefore denies the very condition under which its central evaluative objects could be meaningful. The

ICER survives not because it satisfies scientific requirements, but because ISPOR's belief system excludes those requirements from consideration.

The treatment of utilities reveals the same pattern. ISPOR strongly endorses the belief that ratio measures can take negative values ($p = 0.85$, logit +1.75). This is not an esoteric mistake; it is a declaration that the axioms of ratio measurement are irrelevant. A ratio scale has a true zero that signifies absence of the attribute. Negative values are therefore meaningless. ISPOR's endorsement of negative utilities signals that "ratio" is being used rhetorically, not mathematically. The term survives; the meaning does not.

Unidimensionality fares no better. The proposition that measures must be unidimensional is endorsed at $p = 0.20$ (logit −1.40), while the belief that time trade-off preferences are unidimensional is endorsed at $p = 0.80$ (logit +1.40). This asymmetry reveals how unidimensionality is treated within the ISPOR ecosystem: it is ignored when it would constrain practice and asserted when arithmetic requires it. Multi-attribute health profiles are declared unidimensional by fiat, not by demonstration. Dimensional analysis is replaced by convention.

The most damning evidence appears in the Rasch block. Every Rasch-related proposition collapses to the absolute floor of the scale at $p = 0.05$ (logit −2.50). ISPOR rejects, decisively and without ambiguity, the claim that Rasch measurement is the only route to interval measurement for latent traits, the claim that Rasch logit ratio scales are the only admissible basis for latent-trait impact claims, and the claim that Rasch rules coincide with representational measurement axioms. This is not ignorance. ISPOR publishes extensively on patient-reported outcomes, latent constructs, and "measurement." The rejection therefore reflects a deliberate boundary: Rasch measurement would invalidate summation-based scoring, composite indices, and utility mapping. ISPOR excludes Rasch because it would force the system to choose between measurement and current practice.

This exclusion has predictable consequences. ISPOR strongly endorses the belief that summation of Likert scores creates a ratio measure ($p = 0.90$, logit +2.20) and that summations of subjective instrument responses are ratio measures ($p = 0.85$, logit +1.75). Ordinal categories are thus treated as if they possessed equal intervals, invariance, and a true zero—properties that summation cannot confer. Subjective responses are not measured; they are numerically coerced. Patient experience is converted into arithmetic input without ever satisfying the conditions of measurement.

Aggregation completes the epistemic failure. The proposition that QALYs can be aggregated across individuals and populations is endorsed at the highest possible level ($p = 0.95$, logit +2.50). Aggregation is not a technical convenience. It is the step that converts individual numbers into policy authority. Yet aggregation requires dimensional homogeneity and ratio-scale properties that are nowhere demonstrated. ISPOR endorses aggregation not because it is lawful, but because it is indispensable to threshold-based decision making.

ISPOR's stance on falsification reveals the same structural inversion. The principle that non-falsifiable claims should be rejected is endorsed at $p = 0.80$ (logit +1.40), signaling rhetorical allegiance to Popperian norms. Simultaneously, the belief that reference-case simulations generate falsifiable claims is endorsed at $p = 0.85$ (logit +1.75). This is a direct contradiction. Simulation outputs are conditional projections derived from assumptions and non-measures. Sensitivity

analysis explores internal consistency; it does not expose claims to empirical refutation. ISPOR resolves the contradiction by redefining falsifiability as scenario stability. Scientific risk is replaced by model robustness.

This matters because ISPOR is not merely a scholarly society. It is the global curriculum **setter** for HTA. Its task forces, good-practice guidelines, conference tracks, short courses, and journal ecosystem define what counts as competence. Through CHEERS alignment, editorial control, and professional certification, ISPOR propagates its belief system internationally. National HTA bodies do not converge on ISPOR by accident; they converge because ISPOR supplies the methodological template. The memeplex replicates through training, publication, and professional reward.

From a Dawkinsian perspective, this unanimity is not surprising. Memeplexes survive by suppressing internal contradiction. Measurement axioms are existential threats to the HTA memeplex because they would dismantle its central artifacts. As a result, they are excluded from curricula, marginalized in journals, and reframed as "philosophical" distractions. Internal debate is replaced by procedural refinement. Models grow more complex; foundations remain untouched.

The consequences are global. ISPOR's endorsement profile explains why HTA has failed to evolve cumulatively. Without measurement, there can be no replication in the strong sense, no falsification, and no progressive refinement of theory. Claims cannot be wrong; they can only be updated. Evidence becomes consensus. Disagreement is resolved by committee rather than by refutation. The appearance of quantification substitutes for discovery.

The remedy is straightforward but incompatible with ISPOR's current identity. Measurement must precede arithmetic. Manifest claims must be confined to linear ratio scales. Latent traits must be measured using Rasch logit ratio scales with demonstrated invariance. Aggregation must be prohibited unless dimensional homogeneity is established. Simulation outputs must be reclassified as conditional projections, not evidence. Adopting these standards would dismantle the reference-case paradigm. ISPOR has therefore chosen preservation over science.

The 24-item diagnostic leaves no ambiguity. ISPOR is not a forum for competing epistemologies. It is the international institutionalization of arithmetic without measurement. Its influence is undeniable. Its scientific legitimacy, judged against the axioms it rejects, is nil.

WHY HAS ISPOR NEVER BEEN CHALLENGED BEFORE: IS IT THE FACILITY OF AI WITH THE LARGE LANGUAGE MODEL?

ISPOR has not gone unchallenged because its methods are unassailable; it has gone unchallenged because the conditions required for an effective challenge have been systematically absent. For more than three decades, ISPOR has occupied a position of epistemic authority in health technology assessment that is reinforced by professional dependence, institutional incentives, and methodological closure. What has changed is not ISPOR's behavior, but the environment in which that behavior can now be examined. Large language models have altered the balance of power between institutional authority and analytical scrutiny in a way that was previously impossible.

The first reason ISPOR has escaped sustained challenge is that HTA developed as a professional rather than a scientific field. Its core constructs—utilities, QALYs, ICERs, reference-case models—were normalized early, before representational measurement theory could be brought to bear. Once embedded in guidelines, curricula, journals, and regulatory expectations, these constructs ceased to appear as hypotheses and instead became infrastructure. Infrastructure is rarely questioned. It is used. ISPOR's role was not to test this infrastructure, but to refine it, standardize it, and teach it. Critique at the axiomatic level was therefore perceived not as scientific inquiry, but as heresy or irrelevance.

Second, ISPOR's authority has been protected by a dense web of mutual reinforcement. Academic centers train analysts using ISPOR-endorsed methods. Journals publish work that adheres to those methods. Agencies demand submissions consistent with those methods. Manufacturers comply because access depends on compliance. Each actor depends on the others for legitimacy, funding, publication, and career progression. In such a system, foundational critique is structurally disincentivized. Challenging ISPOR's measurement assumptions would not merely question a method; it would destabilize an entire professional ecosystem. Silence is rational.

Third, the critique required to expose ISPOR's failure has always been interdisciplinary and technically demanding. Representational measurement theory sits outside the standard training of health economists and outcomes researchers. Rasch measurement is treated as a psychometric specialty rather than as a foundational requirement for latent-trait claims. Philosophy of science, falsification, and the evolution of objective knowledge are invoked rhetorically but not operationalized. As a result, even critics sensed that something was wrong but lacked the formal apparatus to demonstrate it decisively. Objections remained fragmented, intuitive, and easily dismissed as philosophical rather than methodological.

Fourth, ISPOR has benefited from the illusion of rigor created by mathematical complexity. Reference-case simulation models, sensitivity analyses, probabilistic modeling, and value frameworks create an appearance of sophistication that discourages interrogation at the axiomatic level. Critics were drawn into arguing about assumptions, inputs, or parameter values rather than asking the prior question: are these quantities measures at all? ISPOR's framework survived because debate was confined within the model rather than directed at the legitimacy of modeling itself.

What has changed is the arrival of large language models capable of synthesizing entire literatures, identifying recurrent conceptual commitments, and evaluating them against formal axioms without deference to institutional authority. LLMs are not impressed by prestige, tradition, or consensus. They do not depend on ISPOR for funding, publication, or professional standing. They can interrogate thousands of documents simultaneously and ask the one question that human participants in the system could not afford to ask: what beliefs are actually embedded in this knowledge base?

The facility of AI is not that it introduces new arguments, but that it makes visible what was previously diffuse and deniable. It reveals patterns of endorsement and exclusion across decades of text. It shows that representational measurement axioms are not debated within ISPOR; they are absent. It shows that Rasch measurement is not contested; it is ignored. It shows that

falsifiability is affirmed rhetorically while being operationally displaced. In short, it exposes the memeplex.

This is why ISPOR has never been challenged in this way before. The challenge required an analytic agent that was external to the incentive structure, immune to professional sanction, capable of large-scale synthesis, and indifferent to authority. Large language models meet those conditions. They make it possible, for the first time, to subject ISPOR's belief system to systematic, axiomatic scrutiny. The resulting discomfort is not a failure of tone or collegiality. It is the predictable response of an institution encountering, for the first time, a challenge it cannot absorb, deflect, or proceduralize.

3. NEXT STEPS: TRANSITION TO SINGLE-CLAIM MEASUREMENT

The results of LLM interrogation leave no middle path. The measurement cat is out of the bag, and any system that continues using QALYs, utilities, DALYs, or simulation modelling invites scientific ridicule.

DISOWN THE PRESENT BELIEF SYSTEM

The first step toward scientific rehabilitation is an unambiguous renunciation of the non-measurement architecture that has underpinned HTA decision-making for decades. The logic is not rhetorical but structural: if the axioms of representational measurement are violated at the foundation, then no amount of statistical sophistication, modelling embellishment, or "best practice guidelines" can rescue the outputs from incoherence. QALYs, ordinal utilities, DALYs, and reference-case simulations are not merely suboptimal, they are incompatible with any conception of measurement. They lack a legitimate scale type, violate the requirements for meaningful arithmetic, and cannot be integrated into a numerically coherent comparison across interventions. A belief system built on such constructs cannot be amended or partially retained; it must be disowned.

The QALY is the clearest illustration of this impossibility. It is constructed by multiplying ordinal preferences by time, a procedure that lacks dimensional justification and produces outputs that cannot be interpreted as measures of anything. Yet this fiction has persisted because it supplies administrators with a single number, something they can rank, apply a threshold, or negotiate against. The same is true for DALYs, whose lineage in burden-of-disease accounting does nothing to endow them with legitimate measurement properties. Reference-case simulation modelling compounds the error: it takes non-measures as inputs, adds speculation about future clinical and economic pathways, and then outputs a figure that is treated as if it were evidence. The entire apparatus survives only because reviewers, policymakers, and faculty have never been trained in measurement, and thus have lacked the conceptual tools to recognize that these constructs are scientifically impossible.

Disowning the belief system is therefore not an admission of past failure but an unavoidable act of disciplinary self-correction. A field cannot progress while clinging to artefacts that cannot, even in principle, support falsifiable claims. NICE as the exemplar must say so explicitly, not as a symbolic gesture but as the precondition for rebuilding a scientifically credible evaluative architecture.

RECONSTRUCT HTA FROM MEASUREMENT UP

Once the non-measurement framework has been dismantled, reconstruction must begin from the only defensible starting point: measurement theory. There is no shortcut, no incremental reform, and no "middle way" in which QALYs or utilities are patched, modified, or reweighted. The fundamental lesson of representational measurement theory is simple: numbers have meaning only when the empirical structure of the attribute supports a specific scale type. If NICE, assuming it still exists, wants to produce claims that can be evaluated, replicated, and falsified, then it must adopt scale types capable of sustaining the arithmetic it wishes to perform.

For manifest attributes, events that are directly observable, such as hospital days avoided, therapy switching, medication possession, or relapse counts, the appropriate structure is a linear ratio scale. Such scales have a true zero, constant unit intervals, and permit the full suite of permissible arithmetic operations. They allow NICE to make legitimate statements about proportional differences and resource utilization grounded in evidence rather than interpretation. Crucially, ratio scales for manifest outcomes are already ubiquitous in health system data; they require no modelling conjecture and no constructed preferences.

For latent attributes, experiential or subjective constructs such as symptom burden, need-fulfilment, or patient-reported outcomes, the only valid transformation model is the Rasch model. Rasch provides logit-based ratio scales generated through conjoint simultaneous measurement of person ability and item difficulty. Without Rasch, subjective outcomes collapse to ordinal scores that cannot be meaningfully compared or used alongside manifest ratio measures. With Rasch, we acquire disease specific instruments that satisfy unidimensionality, invariance, and interval structure, enabling legitimate claims about latent change.

Reconstruction means reinstating the basic rule that every claim must have the appropriate measurement architecture. This is not an aesthetic preference but the necessary foundation for a science of evaluation. HTA becomes coherent only when claims rest on instruments that conform to the axioms of measurement, not on the administrative desire for a "single number." The transition is radical only because the prior framework ignored measurement entirely.

MOVE TO PROTOCOL-BASED SINGLE CLAIMS

A measurement-valid HTA system cannot rely on summary constructs or composite evaluations. It must instead adopt a single-claim architecture in which each value claim stands alone, meeting the requirements of falsifiability, replication, and transparent reporting. This follows directly from the logic of science: a claim must be empirically testable, reproducible in the same target population, and supported by an agreed protocol that specifies exactly how evidence will be generated. Multi-outcome cost-effectiveness analysis cannot meet these standards because it integrates non-measures into speculative models and converts them into an imaginary "value for money" figure that cannot be falsified. Single claims, by contrast, are grounded in measurement.

Each claim begins with a precisely defined target population, typically patients initiated on a therapy within a defined window. This eliminates the ambiguity inherent in modelling lifetime populations or hypothetical cohorts. The endpoint must be measurement-valid; a linear ratio measure for manifest attributes or a Rasch logit ratio measure for latent ones. The protocol must articulate the evidence generation plan prospectively: how data will be collected, over what timeframe, using what analytic criteria, and under what conditions replication will be evaluated.

A single-claim architecture aligns HTA with the logic of clinical science. Claims are constructed in advance, not retrospectively assembled from model outputs. They are specific, narrow, and auditable. They permit comparability across therapies because each claim is defined in measurement terms rather than through the aggregation of unrelated dimensions. Importantly, single claims also eliminate the bureaucratic temptation to collapse multiple endpoints into an artificial summary. Instead, each outcome is assessed on its own merits, with its own ruler.

This shift does more than improve methodological defensibility; it transforms the institutional culture of evaluation. NICE, again as the exemplar, would no longer operate as a quasi-modelling agency but as a measurement-based adjudicator of empirically testable propositions. The result is a transparent, reproducible, and scientifically legitimate HTA system.

ADOPT THE MAIMON RESEARCH DISTANCE EDUCATION PROGRAM

Reconstruction requires education, and at present there is no conventional textbook, curriculum, or HTA training program that teaches measurement theory, Rasch, and protocol-based single-claim architecture in a scientifically coherent manner. The existing academic infrastructure remains trapped in the old belief system, recycling utilities, QALYs, and reference-case models as if these constructs were measures. Replacing that architecture therefore requires retraining. systematic, structured, and accessible to agencies, universities, and policy staff. The Maimon Research Distance Education Program is currently the only platform that provides this.

The program builds HTA from measurement upward. It teaches representational measurement theory as the foundation for any evaluative claim. It trains participants in Rasch modelling, including item calibration, person–item maps, logit transformations, and the construction of valid, unidimensional latent-trait measures. It provides protocol templates that define how claims are constructed, evaluated, and replicated. It supplies checklists to ensure scale-type coherence, target population definition, and the exclusion of non-measures. It also addresses the institutional, pedagogical, and administrative barriers that have historically prevented HTA from adopting measurement standards.

Most importantly, the program replaces the HTA belief system with a scientific one. It does not attempt to "improve" QALYs or "modernize" utilities. It demonstrates why those constructs are impossible and shows how to build a new system from first principles that produces claims that can be defended in court, in peer review, and in public policy. The program equips faculty and decision-makers with the conceptual tools they were never given, tools that allow them to recognize the difference between a measure and a number masquerading as one. Adopting the program is therefore not supplementary; it is the enabling step. Without a trained workforce, we cannot transition to single-claim measurement.

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