MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

THE *EUROPEAN JOURNAL OF HEALTH ECONOMICS*: INSTITUTIONALIZING FALSE MEASUREMENT

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

LOGIT WORKING PAPER No 475 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

FOREWORD

HEALTH TECHNOLOGY ASSESSMENT: A GLOBAL SYSTEM OF NON-MEASUREMENT

The European Journal of Health Economics presents itself as one of the leading international forums for the analysis of health care financing, health policy, and economic evaluation of medical technologies. Established to support the development and application of economic reasoning in health care decision-making, the journal positions itself at the intersection of academic research and policy implementation. Its stated mission is to publish rigorous theoretical and empirical work that informs resource allocation, reimbursement, and pricing decisions across national health systems. In doing so, it contributes directly to the intellectual infrastructure of health technology assessment, providing the analytical tools and evaluative frameworks used to assess therapeutic value, cost-effectiveness, and comparative clinical impact.

The journal's influence extends well beyond academic discourse. Its articles are routinely cited in national HTA guidelines, reimbursement submissions, and policy deliberations across Europe and internationally. By publishing methodological contributions on cost-effectiveness analysis, utility measurement, and simulation modeling, the journal plays a formative role in defining what is accepted as valid evidence in economic evaluation. It therefore functions not merely as a passive repository of research, but as an active participant in shaping the evaluative standards that govern therapy access, pricing negotiations, and resource allocation. The journal's authority rests on its claim to provide scientifically grounded quantitative methods capable of supporting rational and accountable decision-making in modern health care systems.

The purpose of this evaluation is to determine whether the quantitative claims published within the European Journal of Health Economics rest upon constructs that satisfy the axioms of representational measurement. The journal asserts a central role in informing therapy evaluation, pricing, and resource allocation decisions, yet such decisions require that numerical claims possess demonstrable measurement properties. Arithmetic operations, including multiplication, aggregation, and ratio comparison, are admissible only when applied to quantities measured on appropriate scales. This interrogation therefore assesses whether the journal's knowledge base recognizes and enforces these prerequisites, or whether it relies on composite scores and preference-weighted indices that lack lawful scale structure. This assessment is important because without valid measurement, quantitative claims cannot support falsification, replication, or the evolution of objective knowledge, and therefore cannot serve as a scientifically credible basis for health care decision-making.

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

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 EUROPEAN JOURNAL OF HEALTH ECONOMICS KNOWLEDGE BASE

The knowledge base of the European Journal of Health Economics is defined by its explicit commitment to advancing economic evaluation as a tool for informing health care decision making. The journal positions itself at the intersection of economics, clinical research, and public policy, providing a platform for studies that assess the costs and outcomes of medical interventions and their implications for resource allocation. Its published work encompasses cost-effectiveness analyses, cost-utility studies, burden-of-disease assessments, and methodological investigations into the measurement of health-related quality of life. These contributions are presented as quantitative evaluations intended to support comparative judgments regarding the value of competing therapies, guide reimbursement decisions, and inform national and international health policy frameworks.

At the center of this evaluative architecture is the routine use of composite outcome constructs, particularly quality-adjusted life years, which combine duration of life with preference-based assessments of health status. These preference weights are typically derived from multiattribute instruments such as the EQ-5D, which describe health across multiple domains and assign numerical values to combinations of domain responses. These values are then treated as if they represent measurable quantities that can be multiplied by time and aggregated across individuals to produce summary estimates of therapeutic impact. The journal’s knowledge base therefore presumes that multidimensional health descriptions can be converted into single numerical indices capable of supporting arithmetic operations and comparative evaluation across interventions, disease areas, and populations.

Simulation modeling plays a central role in operationalizing this framework. Many published studies employ decision-analytic models to project long-term costs and outcomes beyond the duration of clinical trials. These models integrate clinical inputs, epidemiological assumptions, and preference-based utility scores to generate estimates of incremental cost per quality-adjusted life year gained. The resulting outputs are presented as quantitative indicators of value, capable of informing reimbursement thresholds and resource allocation decisions. This modeling architecture allows analysts to produce internally consistent numerical projections even when direct empirical observation of long-term outcomes is not available.

The journal’s knowledge base also incorporates extensive discussion of methodological refinements intended to improve the precision and policy relevance of economic evaluation. These include improved methods for utility elicitation, adjustments for uncertainty through probabilistic sensitivity analysis, and refinements in model structure and parameter estimation. Such developments are presented as strengthening the analytical framework by increasing transparency, consistency, and comparability across studies. The emphasis is on methodological sophistication, computational rigor, and alignment with established health technology assessment guidelines adopted by national and international agencies.

At the same time, the numerical constructs employed within this framework originate from scoring systems based on subjective assessments of health states rather than from demonstrated measurement structures satisfying representational measurement requirements. Composite utility indices are generated through valuation exercises that assign numerical values to multidimensional health descriptions, and these values are subsequently treated as if they possess the properties required for arithmetic manipulation. The journal's knowledge base therefore operates within an evaluative paradigm in which scoring algorithms and preference aggregation serve as the foundation for quantitative claims regarding therapeutic impact.

This framework has achieved widespread institutional acceptance and is presented within the journal as the standard method for quantitative evaluation in health economics. Its outputs are used to support policy recommendations, inform reimbursement decisions, and guide resource allocation across health systems. The journal's role is therefore not merely descriptive but constitutive. It contributes to the reproduction and reinforcement of the methodological conventions that define contemporary health economic evaluation. Its published analyses shape how therapeutic value is conceptualized, quantified, and communicated within the broader health technology assessment environment.

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.*

CATEGORICAL PROBABILITIES

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: EUROPEAN JOURNAL OF HEALTH ECONOMICS

Table 1 presents, the endorsement 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: EUROPEAN JOURNAL OF HEALTH ECONOMICS: THE ABSENCE OF MEASUREMENT

The EJHE presents itself as both "highly scientific" and "practice oriented," explicitly aiming to cover pharmacoeconomics, pricing and reimbursement systems, and "quality-of-life-studies" alongside core methods. That declared scope matters because it locates the journal at the exact hinge-point where measurement discipline should be non-negotiable: if you publish claims intended to steer reimbursement, pricing, and access, then the first obligation is that the numbers used to justify those decisions are actually measures. The table shows the opposite. The EJHE knowledge base reproduces, with European polish and methodological self-confidence, the same epistemic inversion you have documented elsewhere: arithmetic is treated as the starting point, while the axioms that make arithmetic meaningful are treated as optional, peripheral, or simply nonexistent.

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	0.15	-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	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.90	+2.20
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.85	+1.75
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.55	+0.50
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20

THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.30	-0.95
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	0.40	-0.45
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

Start with the foundational "true" propositions. "Measurement precedes arithmetic" sits at 0.10 (−2.20). "Meeting the axioms of representational measurement is required for arithmetic" is also 0.10 (−2.20). These are not subtle or controversial claims. They are definitional constraints on what it means to quantify anything. When a knowledge base assigns them near-floor endorsement, it is not expressing a nuanced position; it is revealing that the discipline, as practiced and normalized within the journal's publication ecology, does not recognize these constraints as binding. Put plainly: the EJHE corpus behaves as if one may proceed directly to calculation, thresholds, ratios, incremental comparisons, without first demonstrating that the entities being manipulated possess the scale properties required by those operations.

That inversion becomes sharper when paired with the journal's posture on multiplication. "Multiplication requires a ratio measure" sits at 0.10 (−2.20), yet the false claim "the QALY is a ratio measure" sits at 0.95 (+2.50), and "QALYs can be aggregated" sits at 0.95 (+2.50). This is the signature pattern of the HTA memeplex: the precondition is denied while the prohibited operation is celebrated. The journal's knowledge base behaves as if the problem of admissible arithmetic has already been solved; indeed, solved so thoroughly that cost-per-QALY ratios and their aggregation can be treated as routine objects of scientific discourse. The table says that, in practice, EJHE treats the QALY as if it were a legitimate quantity for multiplication, addition, and cross-therapy comparison, while simultaneously failing to endorse the minimal rule that would make those operations lawful.

The same pattern is visible in the handling of scale type. "Interval measures lack a true zero" is only 0.20 (−1.40), "measures must be unidimensional" is 0.15 (−1.75), and the false claims that keep utility arithmetic alive—"EQ-5D preference algorithms create interval measures," "summations of subjective instrument responses are ratio measures," "summation of Likert scores creates a ratio measure"—cluster at 0.90 (+2.20) or higher. This is not an accidental misunderstanding; it is an operational doctrine. The journal's knowledge base systematically elevates scoring conventions into measurement claims, then treats those claims as licenses for

arithmetic. Once you accept that a preference algorithm "creates" an interval measure, you have granted permission for subtraction, averaging, and parametric modeling; once you accept that summed ordinal responses are "ratio," you have granted permission for multiplication and division. The table shows that EJHE's knowledge base repeatedly grants those permissions through false endorsements.

Dimensional homogeneity is the quiet killer here, and the journal's profile reveals it in a particularly instructive way. "The QALY is a dimensionally homogeneous measure" is false, yet it sits at 0.85 (+1.75). That is, the knowledge base treats the QALY not merely as usable but as structurally coherent. But a QALY is built by multiplying time (a manifest ratio quantity) by a preference weight derived from multidimensional health state descriptions and valuation conventions. The product is then treated as though it were a homogeneous "amount" of something called quality-adjusted life. Even if one bracketed, purely for argument, the ordinal nature of preference weights, the product still fails dimensional homogeneity: it is not "life" in the sense time is life-years, and it is not a measured latent trait in the sense a Rasch-derived logit is a measured trait. The table shows EJHE endorses the fiction of homogeneity as if it were settled science, which is exactly why cost-per-QALY ratios can be treated as natural objects for thresholding and league-table comparison.

Rasch is where the mask comes off entirely. The Rasch core items collapse to the floor: "there are only two classes of measurement linear ratio and Rasch logit ratio" is 0.05 (−2.50); "transforming subjective responses to interval measurement is only possible with Rasch rules" is 0.05 (−2.50); "the Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits" is 0.05 (−2.50); "the Rasch rules for measurement are identical to the axioms of representational measurement" is 0.05 (−2.50). These floor values are decisive because EJHE explicitly positions itself to publish "quality-of-life-studies." If an attribute of quality of life is treated as a latent attribute, where the journal's own scope statement invites that treatment, then there is no measurement path that bypasses Rasch as the unique, necessary and sufficient, set of rules for transforming ordinal scores to ratio measures. A journal that is comfortable publishing quality-of-life claims while simultaneously excluding Rasch at the floor is not occupying a neutral position; it is institutionalizing the substitution of scoring for measurement.

The item "the logit is the natural logarithm of the odds-ratio" at 0.30 (−0.95) adds an important nuance: the issue is not merely that Rasch is absent as a recommended method; the underlying quantitative literacy required to even recognize what a Rasch scale is appears weakly held. That matters because the rhetoric of "utilities," "preference weights," and "indexes" often mimics the rhetoric of measurement while avoiding the obligations that measurement imposes. If a knowledge base does not carry the conceptual furniture to distinguish a logit metric (as a transformation tied to odds structure and invariance requirements) from a scoring algorithm (as a mapping rule), then it becomes easy, almost automatic, to treat any number as if it were a measure provided it can be printed to two decimal places and inserted into a model.

EJHE's partial endorsement of "non-falsifiable claims should be rejected" (0.55, +0.52) might look, to the memeplex reader, like a saving grace. It is not. It functions more like a moral decoration than an operational constraint, because the very next line shows the false proposition "reference case simulations generate falsifiable claims" endorsed at 0.90 (+2.20). This is the familiar

displacement move: the journal can praise falsifiability in principle while treating simulation outputs as if they were exposed to refutation in practice. But simulation outputs are functions of assumptions; they are internally consistent computations, not measurements. When assumptions change, outputs change; when new evidence arrives, models are "updated." That is not falsification. It is recalculation within a closed system. The table shows EJHE's knowledge base has normalized that closure as if it were scientific testing.

The practical implications are straightforward and severe. EJHE's declared "practice oriented" aspiration is precisely what makes this logit profile a duty-of-care problem rather than a technical disagreement. In a pricing and reimbursement environment, journal content is not inert. It shapes methods training, reviewer expectations, and the rhetorical standards that manufacturers and agencies learn to satisfy. If a journal systematically treats non-measures as measures, then it teaches analysts to produce authoritative-looking numbers that cannot, even in principle, support the claims being made for them. That is the operational meaning of the inversion that is documented: instead of measurement grounding arithmetic, arithmetic manufactures the appearance of measurement. Once that is normalized, replication becomes impossible in the scientific sense, because there is no stable unit to replicate. One can repeat a model, repeat a scoring algorithm, repeat a valuation survey. But repeating a convention is not replicating a measurement claim.

This is also why the journal's location in Europe matters. EJHE sits in a policy environment where guideline diffusion is real and where "harmonization" pressures are constant. Its scope explicitly references European guideline introduction as part of the journal's motivating context. When journals in such environments endorse the QALY machinery as if it were measurement-valid, they do more than publish papers; they stabilize a transnational administrative consensus; they reinforce the global HA memeplex of false measurement. That consensus then masquerades as scientific convergence. The logit table shows that what is converging is not measurement discipline but shared permission to do inadmissible arithmetic on composite constructs.

What would a measurement-grounded EJHE look like, given its aims? It would treat pricing and reimbursement claims as hypotheses about therapy impact expressed in ratio-scale units wherever possible: manifest claims in resource units and clinical event counts, and latent claims only where Rasch-based invariant measurement has been established. It would treat multiattribute preference indexes as descriptions, useful perhaps for classificatory or narrative purposes, but not as arithmetic objects. It would require that any claim involving multiplication, division, or aggregation demonstrate ratio properties and dimensional homogeneity. It would distinguish, explicitly and repeatedly, between costs as local prices (non-invariant) and resource units as countable ratio quantities, and it would forbid "cost-effectiveness" ratios built on composite denominators that are not measures. The table shows none of those constraints operate as binding norms in the present knowledge base.

The most damning interpretation is also the simplest. EJHE explicitly invites work on quality-of-life and reimbursement, while its logit profile indicates that the foundational rules of measurement are not part of what counts as publishable quantitative reasoning in that domain. The journal therefore functions as a high-status conduit for the global HTA memeplex: a system that values closure, comparability-by-ritual, and model-based decisiveness over the evolution of objective

knowledge. The journal may be "scientific" in style with peer review, methods sections, robustness checks, but the table indicates it is not scientific in the deeper sense that matters: it does not enforce the preconditions that make quantitative claims falsifiable, replicable, and cumulatively correctable.

In that light, the EJHE logit profile is not merely "bad on measurement," it is diagnostic of institutionalized non-possession. True axioms sit near the floor; false permissions sit near the ceiling. Rasch sits at the absolute floor even while "quality-of-life-studies" are treated as core content. This is exactly the architecture of false measurement as a stabilized belief system: it cannot see measurement-first constraints as legitimate scientific requirements, yet it can see, indeed it can prize, the elaborate arithmetic edifices those constraints would immediately disallow. If EJHE wishes to remain relevant to anyone who still believes HTA should be capable of learning from being wrong, then it has only one path: stop treating the axioms of representational measurement as optional philosophy and start treating them as the entry criteria for quantitative claims. Without that reconstruction, its "practice orientation" becomes something darker: a routinized production line for numbers with administrative authority but without measurement validity.

III. THE TRANSITION TO MEASUREMENT IN HEALTH TECHNOLOGY ASSESSMENT

THE IMPERATIVE OF CHANGE

This analysis has not been undertaken to criticize decisions made by health system, nor to assign responsibility for the analytical frameworks currently used in formulary review. The evidence shows something more fundamental: organizations have been operating within a system that does not permit meaningful evaluation of therapy impact, even when decisions are made carefully, transparently, and in good faith.

The present HTA framework forces health systems to rely on numerical outputs that appear rigorous but cannot be empirically assessed (Table 1). Reference-case models, cost-per-QALY ratios, and composite value claims are presented as decision-support tools, yet they do not satisfy the conditions required for measurement. As a result, committees are asked to deliberate over results that cannot be validated, reproduced, or falsified. This places decision makers in an untenable position: required to choose among therapies without a stable evidentiary foundation.

This is not a failure of expertise, diligence, or clinical judgment. It is a structural failure. The prevailing HTA architecture requires arithmetic before measurement, rather than measurement before arithmetic. Health systems inherit this structure rather than design it. Manufacturers respond to it. Consultants reproduce it. Journals reinforce it. Universities promote it. Over time it has come to appear normal, even inevitable.

Yet the analysis presented in Table 1 demonstrates that this HTA framework cannot support credible falsifiable claims. Where the dependent variable is not a measure, no amount of modeling sophistication can compensate. Uncertainty analysis cannot rescue non-measurement. Transparency cannot repair category error. Consensus cannot convert assumption into evidence.

The consequence is that formulary decisions are based on numerical storytelling rather than testable claims. This undermines confidence, constrains learning, and exposes health systems to growing scrutiny from clinicians, patients, and regulators who expect evidence to mean something more than structured speculation.

The imperative of change therefore does not arise from theory alone. It arises from governance responsibility. A health system cannot sustain long-term stewardship of care if it lacks the ability to distinguish between claims that can be evaluated and claims that cannot. Without that distinction, there is no pathway to improvement; only endless repetition for years to come.

This transition is not about rejecting evidence. It is about restoring evidence to its proper meaning. It requires moving away from composite, model-driven imaginary constructs toward claims that are measurable, unidimensional, and capable of empirical assessment over time. The remainder of this section sets out how that transition can occur in a practical, defensible, and staged manner.

MEANINGFUL THERAPY IMPACT CLAIMS

At the center of the current problem is not data availability, modeling skill, or analytic effort. It is the nature of the claims being advanced. Contemporary HTA has evolved toward increasingly complex frameworks that attempt to compress multiple attributes, clinical effects, patient experience, time, and preferences into single composite outputs. These constructs are then treated as if they were measures. They are not (Table 1).

The complexity of the reference-case framework obscures a simpler truth: meaningful evaluation requires meaningful claims. A claim must state clearly what attribute is being affected, in whom, over what period, and how that attribute is measured. When these conditions are met, evaluation becomes possible. When they are not complexity substitutes for clarity. The current framework is not merely incorrect; it is needlessly elaborate. Reference-case modeling requires dozens of inputs, assumptions, and transformations, yet produces outputs that cannot be empirically verified. Each additional layer of complexity increases opacity while decreasing accountability. Committees are left comparing models rather than assessing outcomes.

In contrast, therapy impact can be expressed through two, and only two, types of legitimate claims. First are claims based on manifest attributes: observable events, durations, or resource units. These include hospitalizations avoided, time to event, days in remission, or resource use. When properly defined and unidimensional, these attributes can be measured on linear ratio scales and evaluated directly.

Second are claims based on latent attributes: symptoms, functioning, need fulfillment, or patient experience. These cannot be observed directly and therefore cannot be scored or summed meaningfully. They require formal measurement through Rasch models to produce invariant logit ratio scales. These two forms of claims are sufficient. They are also far more transparent. Each can be supported by a protocol. Each can be revisited. Each can be reproduced. Most importantly, each can fail. But they cannot be combined. This is the critical distinction. A meaningful claim is one that can be wrong.

Composite constructs such as QALYs do not fail in this sense. They persist regardless of outcome because they are insulated by assumptions. They are recalculated, not refuted. That is why they cannot support learning. The evolution of objective knowledge regarding therapy impact in disease areas is an entirely foreign concept. By re-centering formulary review on single-attribute, measurable claims, health systems regain control of evaluation. Decisions become grounded in observable change rather than modeled narratives. Evidence becomes something that accumulates, rather than something that is re-generated anew for every submission.

THE PATH TO MEANINGFUL MEASUREMENT

Transitioning to meaningful measurement does not require abandoning current processes overnight. It requires reordering them. The essential change is not procedural but conceptual: measurement must become the gatekeeper for arithmetic, not its byproduct.

The first step is formal recognition that not all numerical outputs constitute evidence. Health systems must explicitly distinguish between descriptive analyses and evaluable claims. Numbers that do not meet measurement requirements may inform discussion but cannot anchor decisions.

The second step is restructuring submissions around explicit claims rather than models. Each submission should identify a limited number of therapy impact claims, each defined by attribute, population, timeframe, and comparator. Claims must be unidimensional by design.

Third, each claim must be classified as manifest or latent. This classification determines the admissible measurement standard and prevents inappropriate mixing of scale types.

Fourth, measurement validity must be assessed before any arithmetic is permitted. For manifest claims, this requires confirmation of ratio properties. For latent claims, this requires Rasch-based measurement with demonstrated invariance.

Fifth, claims must be supported by prospective or reproducible protocols. Evidence must be capable of reassessment, not locked within long-horizon simulations designed to frustrate falsification.

Sixth, committees must be supported through targeted training in representational measurement principles, including Rasch fundamentals. Without this capacity, enforcement cannot occur consistently.

Finally, evaluation must be iterative. Claims are not accepted permanently. They are monitored, reproduced, refined, or rejected as evidence accumulates.

These steps do not reduce analytical rigor. They restore it.

TRANSITION REQUIRES TRAINING

A transition to meaningful measurement cannot be achieved through policy alone. It requires a parallel investment in training, because representational measurement theory is not intuitive and has never been part of standard professional education in health technology assessment, pharmacoeconomics, or formulary decision making. For more than forty years, practitioners have been taught to work within frameworks that assume measurement rather than demonstrate it. Reversing that inheritance requires structured learning, not informal exposure.

At the center of this transition is the need to understand why measurement must precede arithmetic. Representational measurement theory establishes the criteria under which numbers can legitimately represent empirical attributes. These criteria are not optional. They determine whether addition, multiplication, aggregation, and comparison are meaningful or merely symbolic. Without this foundation, committees are left evaluating numerical outputs without any principled way to distinguish evidence from numerical storytelling.

Training must therefore begin with scale types and their permissible operations. Linear ratio measurement applies to manifest attributes that possess a true zero and invariant units, such as

time, counts, and resource use. Latent attributes, by contrast, cannot be observed directly and cannot be measured through summation or weighting. They require formal construction through a measurement model capable of producing invariant units. This distinction is the conceptual fulcrum of reform, because it determines which claims are admissible and which are not.

For latent trait claims, Rasch measurement provides the only established framework capable of meeting these requirements. Developed in the mid–twentieth century alongside the foundations of modern measurement theory, the Rasch model was explicitly designed to convert subjective observations into linear logit ratio measures. It enforces unidimensionality, tests item invariance, and produces measures that support meaningful comparison across persons, instruments, and time. These properties are not approximations; they are defining conditions of measurement.

Importantly, Rasch assessment is no longer technically burdensome. Dedicated software platforms developed and refined over more than four decades make Rasch analysis accessible, transparent, and auditable. These programs do not merely generate statistics; they explain why items function or fail, how scales behave, and whether a latent attribute has been successfully measured. Measurement becomes demonstrable rather than assumed.

Maimon Research has developed a two-part training program specifically to support this transition. The first component provides foundational instruction in representational measurement theory, including the historical origins of scale theory, the distinction between manifest and latent attributes, and the criteria that define admissible claims. The second component focuses on application, detailing claim types, protocol design, and the practical use of Rasch methods to support latent trait evaluation.

Together, these programs equip health systems, committees, and analysts with the competence required to enforce measurement standards consistently. Training does not replace judgment; it enables it. Without such preparation, the transition to meaningful measurement cannot be sustained. With it, formulary decision making can finally rest on claims that are not merely numerical, but measurable.

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

For readers who are looking for an introduction to measurement that meets the required standards, Maimon Research has just released two distance education programs. These are:

- Program 1: Numerical Storytelling – Systematic Measurement Failure in HTA.
- Program 2: A New Start in Measurement for HTA, with recommendations for protocol-supported claims for specific objective measures as well as latent constructs and manifested traits.

Each program consists of five modules (approx. 5,500 words each), with extensive questions and answers. Each program is priced at US$65.00. Invitations to participate in these programs will be distributed in the first instance to 8,700 HTA professionals in 40 countries.

More detail on program content and access, including registration and on-line payment, is provided with this link: https://maimonresearch.com/distance-education-programs/

DESIGNED FOR CLOSURE

For those who remain unconvinced that there is any need to abandon a long-standing and widely accepted HTA framework, it is necessary to confront a more fundamental question: why was this system developed and promoted globally in the first place?

The most plausible explanation is administrative rather than scientific. Policy makers were searching for an assessment framework that could be applied under conditions of limited empirical data while still producing a determinate conclusion. Reference-case modeling offered precisely this convenience. By constructing a simulation populated with assumptions, surrogate endpoints, preference weights, and extrapolated time horizons, it became possible to generate a numerical result that could be interpreted as decisive. Once an acceptable cost-effectiveness ratio emerged, the assessment could be declared complete and the pricing decision closed. This structure solved a political and administrative problem. It allowed authorities to claim that decisions were evidence-based without requiring the sustained empirical burden demanded by normal science. There was no requirement to formulate provisional claims and subject them to ongoing falsification. There was no obligation to revisit conclusions as new data emerged. Closure could be achieved at launch, rather than knowledge evolving over the product life cycle.

By contrast, a framework grounded in representational measurement would have imposed a very different obligation. Claims would necessarily be provisional. Measurement would precede arithmetic. Each therapy impact claim would require a defined attribute, a valid scale, a protocol, and the possibility of replication or refutation. Evidence would accumulate rather than conclude. Decisions would remain open to challenge as real-world data emerged. From an administrative standpoint, this was an unreasonable burden. It offered no finality.

The reference-case model avoided this problem entirely. By shifting attention away from whether quantities were measurable and toward whether assumptions were plausible, the framework replaced falsification with acceptability. Debate became internal to the model rather than external to reality. Sensitivity analysis substituted for empirical risk. Arithmetic proceeded without prior demonstration that the objects being manipulated possessed the properties required for arithmetic to be meaningful.

Crucially, this system required no understanding of representational measurement theory. Committees did not need to ask whether utilities were interval or ratio measures, whether latent traits had been measured or merely scored, or whether composite constructs could legitimately be multiplied or aggregated. These questions were never posed because the framework did not require

them to be posed. The absence of measurement standards was not an oversight; it was functionally essential.

Once institutionalized, the framework became self-reinforcing. Training programs taught modeling rather than measurement. Guidelines codified practice rather than axioms. Journals reviewed technique rather than admissibility. Over time, arithmetic without measurement became normalized as "good practice," while challenges grounded in measurement theory were dismissed as theoretical distractions. The result was a global HTA architecture capable of producing numbers, but incapable of producing falsifiable knowledge. Claims could be compared, ranked, and monetized, but not tested in the scientific sense. What evolved was not objective knowledge, but institutional consensus.

This history matters because it explains why the present transition is resisted. Moving to a real measurement framework with single, unidimensional claims does not merely refine existing methods; it dismantles the very mechanism by which closure has been achieved for forty years. It replaces decisiveness with accountability, finality with learning, and numerical plausibility with empirical discipline. Yet that is precisely the transition now required. A system that avoids measurement in order to secure closure cannot support scientific evaluation, cumulative knowledge, or long-term stewardship of healthcare resources. The choice is therefore unavoidable: continue with a framework designed to end debate, or adopt one designed to discover the truth.

Anything else is not assessment at all, but the ritualized manipulation of numbers detached from measurement, falsification, and scientific accountability.

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