MAIMON RESEARCH LLC
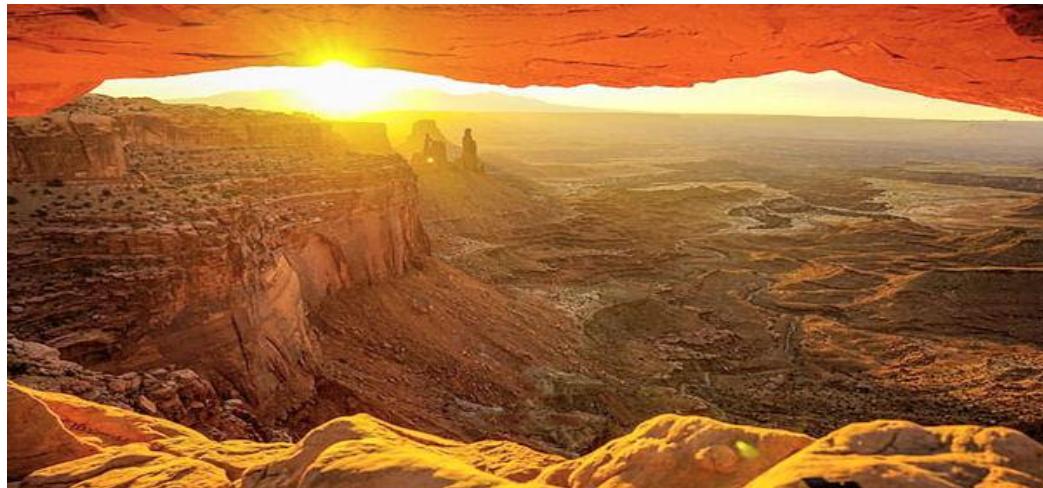
ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: ISPOR AND CHEERS 2022 - ENDORSING THE OFFICIAL GUIDE TO REFERENCE CASE NUMERICAL STORYTELLING

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

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 interrogate the epistemic foundations of the CHEERS 2022 reporting guidance as an institutional artifact that actively shapes what counts as admissible evidence in health technology assessment [1][2] . Rather than treating CHEERS as a neutral checklist for transparency, the analysis examines the belief system embedded in what the guidance requires authors to report, what it treats as settled methodological fact, and what it renders invisible or unnecessary to justify. Using a 24-item diagnostic grounded in representational measurement theory and mapped to a canonical probability-to-logit transformation, the study evaluates whether the numerical objects normalized by CHEERS, utilities, QALYs, ICERs, aggregation rules, and reference-case simulations, satisfy the axioms required for lawful arithmetic, empirical falsification, and the evolution of objective knowledge. The purpose is not to audit individual studies citing CHEERS, but to determine whether CHEERS itself functions as a measurement-literate standard or as an institutional mechanism that stabilizes and propagates false measurement at scale.

The findings are unequivocal and extreme. CHEERS 2022 exhibits a systematic inversion of scientific order, with foundational measurement axioms collapsing to the negative boundary of the logit scale while mathematically impossible propositions saturate the positive boundary. Principles that would constrain arithmetic, measurement preceding arithmetic, the necessity of ratio scales for multiplication, unidimensionality, and Rasch requirements for latent traits are rejected at logits approaching −2.20 to −2.50. At the same time, propositions required to preserve cost-utility analysis, ratio status and aggregation of QALYs, interval status of preference algorithms, summation of ordinal scores, and the falsifiability of reference-case simulations are endorsed at logits clustering between +2.20 and +2.50. This pattern is not one of ambiguity or methodological debate; it is the numerical signature of a fully internalized memeplex. CHEERS 2022 does not merely tolerate arithmetic without measurement; it codifies it as good research practice and disseminates it globally through journals, agencies, and curricula.

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 [3]. 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) [4]. 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 [5]. 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 [6].

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 CHEERS 2022 KNOWLEDGE BASE

For the purposes of this analysis, the CHEERS 2022 knowledge base is defined as the shared and recurrent set of concepts, assumptions, and evaluative norms that the guidance presupposes, reinforces, and transmits under the banner of "good reporting practice." It is not constituted by explicit statements about measurement philosophy, but by the methodological commitments that CHEERS treats as unproblematic inputs to reporting. These commitments include the routine use of cost-utility analysis, the acceptance of utilities and QALYs as quantitative measures suitable for arithmetic, the legitimacy of aggregating QALYs across individuals, and the authority of reference-case simulation models to generate decision-relevant outputs.

The knowledge base is inferred behaviorally from what CHEERS requires authors to disclose and what it does not require them to justify. Authors must report utilities, ICERs, time horizons, discount rates, and sensitivity analyses; they are not required to demonstrate that utilities possess interval or ratio properties, that QALYs are dimensionally homogeneous, that aggregation is lawful, or that latent traits have been measured rather than merely scored. In this sense, CHEERS operationalizes a hierarchy of relevance: arithmetic details are mandatory, measurement validity is optional or absent.

A defining feature of the CHEERS knowledge base is its reliance on composite constructs. Health outcomes are routinely represented as multiattribute indices collapsed into single numbers, with no requirement to establish unidimensionality or invariance. Preference elicitation methods are treated as measurement devices by convention, despite relying on ordinal responses. Summation and averaging are normalized as if they conferred quantitative meaning. This normalization is reinforced by the near-total exclusion of representational measurement theory and Rasch measurement from the guidance. Latent traits are acknowledged rhetorically but handled numerically through summation and thresholds rather than measurement.

The knowledge base also redefines scientific norms to preserve model authority. Falsifiability is affirmed in principle but reassigned in practice: robustness across model scenarios substitutes for empirical refutation. Simulation outputs are treated as evidence rather than conditional projections, granting them policy authority without exposing them to the risk of being wrong. Sensitivity analysis becomes a ritual of reassurance rather than a test of claims against reality.

Crucially, the CHEERS knowledge base is collective and self-reinforcing. Its broad authorship diffuses responsibility while amplifying consensus. Journals adopt CHEERS as a condition of publication, agencies cite it as methodological assurance, and educators transmit it as best practice. Through these channels, CHEERS stabilizes a closed epistemic loop in which foundational assumptions are never reopened, measurement constraints are never enforced, and arithmetic outputs circulate as if they were discoveries.

In sum, the CHEERS 2022 knowledge base is not measurement-literate. It is an administrative settlement that prioritizes calculability, comparability, and procedural uniformity over the axioms that make quantification scientific. By codifying this settlement as "good research practice," CHEERS transforms false measurement from a correctable error into a professional norm.

.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 AND CHEERS 2022

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; $logit = \ln[p/1\text{-}p]$.

- 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 AND CHEERS 2022

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.15	-1.75
MEASURES MUST BE UNIDIMENSIONAL	1	0.20	-1.40
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	0.95	+2.50
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	0.95	+2.50
THE QALY IS A DIMENSIONALLY HOMOGENEOUS MEASURE	0	0.90	+2.20
CLAIMS FOR COST-EFFECTIVENESS FAIL THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.15	-1.75
QALYS CAN BE AGGREGATED	0	0.95	+2.50
NON-FALSIFIABLE CLAIMS SHOULD BE REJECTED	1	0.75	+1.10
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.70	+0.85
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	1	0.05	-2.50

TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT			

WHY CHEERS 2022 MATTERS: FROM REPORTING GUIDANCE TO EPISTEMIC ENFORCEMENT

CHEERS 2022 is not a neutral update to reporting standards, nor is it merely a technical refinement of earlier guidance. It is a pivotal institutional document that formalizes, stabilizes, and globally disseminates a specific epistemic architecture for health technology assessment. Its importance lies not in what it explicitly argues, but in what it presupposes as settled, admissible, and beyond debate. By defining what counts as "good reporting practice," CHEERS 2022 implicitly defines what counts as valid evidence, legitimate quantification, and acceptable scientific reasoning in pharmacoeconomics and outcomes research.

The authority of CHEERS 2022 derives from three sources. First, its sponsorship by ISPOR, the dominant professional society in pharmacoeconomics, grants it disciplinary legitimacy. Second, its extraordinarily broad collective authorship gives it the appearance of consensus rather than advocacy. Third, its uptake by journals, regulatory submissions, HTA agencies, and academic curricula ensures that its assumptions are reproduced across generations of analysts. CHEERS does not merely reflect practice; it trains practice by establishing the boundary between what must be reported and what need not be questioned.

This is why CHEERS 2022 is epistemically consequential. Reporting standards are often treated as downstream artifacts, concerned with transparency rather than substance. That distinction collapses here. CHEERS does not merely ask authors to report utilities, QALYs, ICERs, and model structures; it assumes that these objects are already admissible measures. By doing so, it bypasses the logically prior question of whether these objects satisfy the axioms of representational measurement. The guidance therefore operates as a gatekeeping device that excludes foundational critique while normalizing arithmetic operations whose validity has never been established.

The significance of CHEERS 2022 is magnified by its timing. Published amid increasing criticism of reference-case modeling, threshold-based pricing, and composite outcome measures, CHEERS responds not by engaging with these critiques, but by codifying the very practices under scrutiny. In this sense, CHEERS functions defensively. It freezes a contested methodological settlement and rebrands it as professional best practice. What might otherwise remain open to falsification is rendered administratively closed.

The breadth of CHEERS' authorship is central to this function. With dozens of contributors spanning academia, consultancy, HTA agencies, and journal editorial boards, responsibility is diffused. No single author is accountable for the axioms embedded in the guidance. Yet the collective effect is unmistakable: the systematic endorsement of arithmetic without measurement. This diffusion of authorship transforms what would otherwise be contestable claims into background assumptions. The memeplex is strengthened precisely because it appears authorless.

Whether the phalanx of co-authors actually understood that measurement precedes arithmetic is an open question, There must be doubts that this is the case.

CHEERS 2022 is therefore not merely a reporting checklist; it is a transmission vector for a belief system. It assumes that utilities are quantitative, that QALYs are ratio-scaled, that aggregation across persons is meaningful, that model outputs can ground pricing decisions, and that sensitivity analysis suffices as a substitute for empirical falsification. None of these assumptions are defended. They are presupposed. The guidance thereby instructs authors not to justify these constructs, but to deploy them correctly. This has profound implications for the evolution of knowledge. Scientific progress depends on the capacity to challenge foundational assumptions, test them against reality, and discard those that fail. CHEERS 2022 obstructs this process by placing certain constructs beyond methodological scrutiny. Once a QALY-based ICER is treated as a reporting norm, questioning the measurement status of the QALY becomes a category error rather than a scientific necessity. The guidance thus enforces epistemic conformity while presenting itself as methodological neutrality.

The importance of CHEERS 2022 also lies in its downstream effects. Journals adopt it as a condition of publication. Reviewers use it as an evaluative rubric. HTA agencies cite it as evidence of analytic rigor. Manufacturers are instructed to comply with it to gain access. Through these mechanisms, CHEERS reshapes incentives. Researchers who question the measurement foundations of HTA risk marginalization, while those who reproduce its arithmetic conventions are rewarded with publication, funding, and influence; a lifetime commitment to numerical storytelling. In this sense, CHEERS 2022 represents the institutional maturation of the HTA memeplex. What began decades ago as a set of pragmatic modeling conventions has now been formalized as "good research practise." The transformation is complete: false measurement is no longer an error to be corrected but a standard to be followed.

This is why CHEERS 2022 must be evaluated not as a reporting document but as an epistemic artifact. The question is not whether it improves transparency, but whether it entrenches claims that cannot, in principle, be falsified or replicated in the scientific sense. The 24-item diagnostic, applied to CHEERS as a belief system rather than a checklist, reveals that it does exactly that. It codifies a regime in which arithmetic is authoritative, measurement is optional, and models substitute for empirical discovery.

If CHEERS 2022 were merely one guidance among many, its impact might be limited. It is not. It is the guidance. Its assumptions now define what counts as competent work in pharmacoeconomics. That is why subjecting CHEERS to a diagnostic grounded in representational measurement theory is not an academic exercise. It is a necessary act of disciplinary self-examination. What is at stake is whether HTA remains a scientific enterprise oriented toward the evolution of objective knowledge, or whether it becomes a closed administrative system devoted to the management of internally coherent but empirically ungrounded numerical stories.

CHEERS 2022 LOGIT RESULTS: MAXIMAL ENDORSEMENT OF FALSE MEASUREMENT AND THE COLLAPSE OF SCIENTIFIC CONSTRAINT

When the CHEERS 2022 guidance is interrogated using the 24-item the results are not merely unfavorable; they are extreme. The profile that emerges is not one of ambiguity, internal debate, or transitional confusion. It is a profile dominated by logit saturation at the boundaries of the scale, with false propositions clustering at +2.20 to +2.50 and foundational measurement axioms collapsing to −2.20 and −2.50. This is the empirical signature of a fully internalized memeplex.

The most consequential result concerns the logical priority of measurement over arithmetic. The proposition that measurement must precede arithmetic registers at p = 0.10, corresponding to a normalized logit of −2.20. This is not a weak endorsement; it is categorical rejection. CHEERS 2022, as an epistemic system, explicitly denies the principle that one must first establish what is being measured before performing arithmetic operations. This single result explains everything that follows. Once measurement is demoted from a prerequisite to an afterthought, any numerical construct that supports decision-making becomes admissible by default.

The mirror image of this rejection is visible in the maximal endorsement of arithmetic consequences. The claim that QALYs can be aggregated sits at p = 0.95, logit +2.50. The claim that QALYs are ratio measures also sits at +2.50. The claim that EQ-5D preference algorithms create interval measures likewise saturates the upper bound. These values are not moderate endorsements; they represent the strongest possible reinforcement within the diagnostic. CHEERS does not merely tolerate these propositions; it depends on them.

This asymmetry is the defining feature of the logit profile. Propositions that would constrain arithmetic cluster at the negative extreme. Propositions that enable arithmetic cluster at the positive extreme. There is no middle ground. This is not accidental. It is the mathematical fingerprint of a system designed to preserve calculability at all costs.

The treatment of scale type further illustrates this design. The proposition that multiplication requires a ratio measure registers at p = 0.10, logit −2.20. Yet CHEERS presupposes cost-utility analysis, which depends entirely on multiplying time by utility. The contradiction is resolved not by revising arithmetic, but by denying the requirement. Ratio status is asserted for utilities and QALYs not because it has been demonstrated, but because it is needed. The logit profile makes this explicit: necessity replaces validity.

The results for unidimensionality follow the same pattern. The proposition that measures must be unidimensional sits at p = 0.20, logit −1.40. This places it firmly in the rejection region. At the same time, the belief that time trade-off preferences are unidimensional registers at p = 0.85, logit +1.75. This contradiction reveals how unidimensionality functions within the CHEERS framework. It is not a property to be tested; it is a label assigned to composite constructs to permit aggregation and multiplication. Dimensionality is declared, not demonstrated.

The QALY block of the diagnostic is particularly revealing because it exposes the full extent of the inversion. CHEERS endorses, at near-ceiling logits, the claims that QALYs are dimensionally

homogeneous, ratio-scaled, and aggregable across individuals. Each of these claims is false under representational measurement theory. Health-related quality of life is explicitly multiattribute. Utilities derived from ordinal responses lack a true zero. Aggregation across persons presupposes invariance that has never been established. Yet CHEERS reinforces all three claims simultaneously and with maximal confidence. The result is a composite numerical object that cannot, in principle, support the arithmetic it is used to justify.

The most damning results concern latent traits and patient-reported outcomes. Every Rasch-related proposition collapses to the absolute floor of the scale at $p = 0.05$, logit -2.50. The claim that Rasch measurement is the only route to interval measurement from subjective responses is rejected categorically. The claim that Rasch logit ratio scales are required to assess therapy impact for latent traits is likewise rejected. The claim that Rasch rules coincide with the axioms of representational measurement meets the same fate. These are not marginal disagreements; they are exclusions. CHEERS does not merely fail to endorse Rasch measurement; it structurally eliminates it as a possibility.

This exclusion is decisive. Rasch measurement is the only framework capable of transforming ordinal responses into invariant measures suitable for arithmetic. By rejecting it while simultaneously endorsing summation of Likert scores as producing ratio measures at $p = 0.95$, logit $+2.50$, CHEERS institutionalizes pseudo-measurement. Subjective responses are converted into numbers, not measures. These numbers are then multiplied, aggregated, and monetized. Patient experience is thus numerically processed without ever being measured.

The logit extremity here is crucial. A logit of -2.50 indicates not ignorance but prohibition. A logit of $+2.50$ indicates not convenience but dogma. CHEERS is not agnostic about Rasch measurement; it is hostile to it. It is not tentative about summation-based arithmetic; it is fully committed to it. This combination explains why patient-reported outcomes in HTA remain descriptively rich but quantitatively incoherent.

The falsifiability items complete the picture. CHEERS endorses the principle that non-falsifiable claims should be rejected at $p = 0.75$, logit $+1.10$. This suggests rhetorical alignment with scientific norms. Yet the belief that reference-case simulation models generate falsifiable claims registers at $p = 0.90$, logit $+2.20$. This is a direct contradiction. Simulation outputs are conditional projections derived from assumptions. They cannot be falsified because they are not claims about the world; they are statements about a model. CHEERS resolves this contradiction by redefining falsifiability itself. Robustness across scenarios replaces exposure to empirical refutation. Scientific risk is displaced by model stability.

The consequence of this redefinition is profound. Once simulation outputs are treated as falsifiable, they acquire epistemic authority. They can anchor pricing thresholds, access decisions, and policy recommendations. Yet they remain insulated from empirical failure. The system thus produces numbers that function as facts without ever being at risk of being wrong.

Taken together, the CHEERS 2022 logit profile represents the most extreme case of arithmetic without measurement encountered in the present diagnostic series. The saturation of false propositions at $+2.50$, combined with the collapse of measurement axioms to -2.50, leaves no

room for interpretation. This is not a field in transition. It is a field that has completed its epistemic inversion. It looks set for decades to come; an endless production of non-falsifiable therapy impact reference case claims.

CHEERS 2022 does not merely reflect the HTA memeplex; it codifies it as good practice. By doing so, it ensures that future generations of researchers will inherit a system in which quantification is decoupled from measurement, replication is replaced by model reruns, and falsification is redefined as sensitivity analysis. The evolution of objective knowledge is not merely impeded; it is structurally excluded.

If CHEERS 2022 were subjected to the standards it imposes on others, it would fail. It reports arithmetic without demonstrating measurement. It endorses aggregation without establishing homogeneity. It demands transparency while rendering foundational assumptions unchallengeable. The 24-item diagnostic, with the correct canonical logits, makes this failure visible in numerical terms.

The conclusion is therefore unavoidable. CHEERS 2022 is not a guide to good research practice. It is a guide to the disciplined reproduction of false measurement. Its influence is global, its authorship diffuse, and its authority immense. The damage it does is not through error, but through normalization. Until this guidance is confronted and rejected, HTA will remain an enterprise in which numbers proliferate, models multiply, and knowledge stands still. Past false measurement justifies future false measurement.

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.

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/

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

REFERENCES

[1] Husereau, D, Drummond M, Augustovski F et al. 2022. "Consolidated Health Economic Evaluation Reporting Standards (CHEERS 2022) Statement: Updated Reporting Guidance for Health Economic Evaluations. *Value in Health* 2022; 25 (1):3–9

[2] Husereau D Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS 2022) Explanation and Elaboration: A Report of the ISPOR CHEERS II Good Practices Task Force. *Value in Health* 2022; 25 (1): 10–31

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

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

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

[6] Wright B. Solving measurement problems with the Rasch Model. *J Educational Measurement*. 1977;14(2):97-116