MAIMON WORKING PAPER No 21 OCTOBER 2025

MAIMON DISTANCE EDUCATION: REQUIRED REPRESENTATIONAL MEASUREMENT STANDARDS FOR A NEW START IN HEALTH TECHNOLOGY ASSESSMENT

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

ABSTRACT

After 50 years of guided measurement failure, health technology assessment (HTA) has two options: first, to declare the subject area bankrupt due to a failure to respect the axioms of representational measurement theory or, to admit failure and propose a new system where there are only two measures, first, a linear ratio scale for manifest resource and utilization claims or, second, Rasch logit ratio scales for latent trait possession claims. These scales or measures apply to specific resource utilization or resource claims. The claims must meet the axiomatic standards for representational or Rasch measurement and be falsifiable. They must be accompanied by a protocol detailing how the claim is to be assessed. To achieve these new standards in HTA, Maimon Research has developed two distance learning programs:

- *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.*

The content for each of these programs is detailed here for each of 5 modules for each program. A link is also provided for those who way wish to purchase these programs (each US$65.00).

INTRODUCTION

For a belief system that has received global acceptance, it is difficult for leaders in health technology assessment to declare after 50 years that the accepted analytical framework for comparative therapy assessment is bankrupt. Unfortunately, bankruptcy was inevitable from day one. From the initial decision to value health stare descriptions, the belief system with utilities, QALY and references case claims for cost effectiveness is nothing more than numerical storytelling. Yet thousands believe in these fairy stories; a global following of numerical nonsense.

The more absurd fact is that while the HTA leadership dug in with health state descriptive valuations, measurement theory had made clear in 1946 with Stevens' contribution to the measurement standards required for arithmetic and in 1971 the formalization of the axioms of representational measurement theory (RMT) [1] [2]. The key parallel development was the Rasch rules for transforming observations to interval latent trait measures (1960) and the demonstration by Wright in 1977 that these rules were consistent with the axioms of representational measurement [3] [4].

DISTANCE EDUCATION PROGRAM ACCESS

Access to these programs is straightforward. There are five modules for each program with questions and answers to support the material resented. Each program is US$65.00 for all five modules. They are accessible through the Maimon Research website https://maimonresearch.com/distance-education-programs/ which gives more details plus direct purchase; for direct access to purchase these programs https://maimonresearch.com/programs-for-purchase/

MAIMON RESEARCH PROGRAM 1

NUMERICAL STORYTELLING: SYSTEMATIC MEASUREMENT FAILURE IN HEALTH TECHNOLOGY ASSESSMENT

HTA can be dismissed in a sentence: it confuses numbers with measures. In science, a string of numerals or symbols becomes a measure only when it preserves the empirical structure of an attribute and obeys the transformation rules set out by RMT (RMT). The RMT axioms: unidimensionality, order, additivity, solvability/cancellation, invariance, are what license arithmetic. Without them, subtraction, averaging, ratios, and products are illegitimate. HTA's main artifacts ignore this gate. Utilities derived from preference tasks lack interval meaning; multiplying them by time to make QALYs violates dimensional homogeneity; disease-specific totals are summed scores that have never earned equal units; cost composites bundle heterogeneous quantities. Rasch modeling shows how latent attributes can be measured lawfully, but HTA never demands it. The result is numerical storytelling dressed as evaluation: outputs that look precise yet have no admissible arithmetic. Until HTA requires evidence that its numbers are measures, its claims are not science but policy theater.

MODULE 1: WHY STEVENS? THE CONTEXT OF 1946

Before Stevens (1946), measurement outside physics lacked firm footing. Physical magnitudes, time, length, mass, implicitly assumed single continua with equal units and true zeros; Campbell's concatenation view tried to justify this by showing that empirical combination preserves additivity. Psychophysics (Weber–Fechner) chased lawful relations for sensations, while Bridgman's operationalism defined concepts by the procedures that produced numbers useful, but no guarantee that numerals preserved structure or licensed arithmetic. Two problems remained: when is any numerical assignment a measure, and how can latent attributes be measured without physical concatenation? Stevens answered the first: he tied scale types (nominal, ordinal, interval, ratio) to their admissible transformations, making the legitimacy of arithmetic explicit; relabeling, order-preserving, positive linear, and similarity transformations, respectively. But he left the second open: he did not supply a method to establish unidimensionality, equal units, and invariance for

latent traits. The post-Stevens program filled that gap: *Foundations of Measurement* formalized representation/uniqueness, and Rasch modeling operationalized latent measurement by constructing logit metrics when data fit.

MODULE 2: AXIOMS OF REPRESENTATIONAL MEASUREMENT THEORY

From 1946 to 1971 the field moved from Stevens' pragmatic typology to a fully axiomatized account of when numbers qualify as measures. Suppes formalized extensive (concatenation) measurement, showing how additivity follows from empirical combination rules [5]. Luce and Tukey's conjoint measurement then identified the cancellation, solvability, and Archimedean conditions under which two or more ordered attributes admit an additive (interval) representation without physical concatenation [6]. This work made precise the representation and uniqueness questions Stevens left open: when does a structure-preserving mapping exist, and what transformations leave a scale's meaning intact? The synthesis arrived with *Foundations of Measurement* (1971) by Krantz, Luce, Suppes, and Tversky A, which proved general representation and uniqueness theorems and tied scale types directly to admissible transformations, invariance, and testable axioms. In parallel, Rasch (1960) provided a probabilistic model that operationalized these ideas for latent traits, yielding logit rulers with specific objectivity when data fit. By 1971, the conceptual and mathematical warrant for lawful measurement was in place.

MODULE 3: SUSTAINED MEASUREMENT FAILURE – THE TIME TRADE OFF (TTO) TECHNIQUE, THE EQ-5D-3L PREFERENCE ALGORITHM AND PREFERENCE UTILITIES

Time trade-off (TTO) starts by asking respondents to trade years of life to "value" verbal health state descriptions. Those raw, preference-laden numbers become the dependent variable in a regression where EQ-5D-3L profiles are encoded with dummy variables for each dimension–level. The fitted "tariff" is then turned into an algorithm: plug any EQ-5D-3L profile into the coefficient recipe, add a constant and any penalty terms, and out comes a single "utility" score. That pathway, from TTO judgments to a tariffed index, produces a convenient number, but not a measure in the sense required by RMT. Unidimensionality is assumed for a multiattribute bundle; additivity across dimensions is imposed without the cancellation and solvability tests that warrant it; invariance fails across elicitation protocols and national tariffs; and protocol features manufacture negative values that violate the Archimedean condition. Because the axioms are not met, the tariffed utilities are context-bound indices. The TTO technique establishes measurement failure in HTA by valuing, incorrectly. composite health-state descriptions rather than a single latent attribute, it violates the unidimensionality requirement at the start. Once that axiom is broken, no regression, tariff, or model can restore lawful arithmetic: the resulting "utilities" are guaranteed non-measures; numerical storytelling dressed up as measurement.

MODULE 4: SUSTAINED MEASUREMENT FAILURE – THE IMPOSSIBLE QALY AND THE CHIMERICAL REFERENCE CASE

The QALY and the reference case are the twin pillars of HTA's orthodoxy, with both failing at the level of measurement. QALYs are built by multiplying chronological time, a true ratio measure, by "utilities" derived from valuing health-state descriptions (e.g., TTO/SG). Those utilities are ordinal preference indices, not interval or ratio measures: they lack unidimensionality, equal units,

invariance, and a defensible zero. Multiplying a non-measure by time violates dimensional homogeneity, so the QALY is not merely imperfect; it is undefined in measurement terms. The reference case institutionalizes this error by mandating cost-per-QALY models and treating their outputs as evidence. What looks like rigor, thresholds, probabilistic sensitivity analysis, elaborate model structure, is precision without meaning. The reference case only supports numerical storytelling. An artifact which fails the axioms of measurement, it cannot support arithmetical operations, and the standards of normal science for falsifiable claims that meet either interval or ratio measurement requirements.

MODULE 5: THE IDENTITY CRISIS OF HTA - NOTHING WITHOUT THE REFERENCE CASE

Health technology assessment faces an existential crisis because it treats numbers as measures without earning that status. The reference case rests on utilities created from preference tasks and then multiplies them by time to form QALYs, a product that violates basic requirements of measurement such as interval spacing, invariance, and dimensional homogeneity. When the denominator is not a measure, the resulting cost-per-QALY ratio has no stable unit; it looks quantitative but carries no lawful arithmetic. This is why "cost-effectiveness" within the reference case is a numerically meaningless claim: the ratio's precision is theatrical, not scientific. Checklists and reporting standards further entrench the illusion by policing presentation while ignoring scale type, so what gets replicated is convention, not knowledge. Because claims generated under the reference case cannot be falsified on measurement grounds, HTA functions as policy ritual rather than science. As long as the reference case remains the decision engine for agencies such as NICE it secures HTA's place as a non-science.

MAIMON RESEARCH PROGRAM 2

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

For fifty years, health technology assessment has practiced numerical storytelling by confusing numbers with measures. To function as a science, HTA must accept the axioms of representational measurement theory: first clarified by Stevens (1946), who tied arithmetic to scale type, and completed by Krantz, Luce, Suppes, and Tversky (1971) with representation and uniqueness theorems. In parallel, Rasch (1960) supplied the probabilistic bridge for latent traits; Wright (1977) showed how ordered responses can be transformed into a logit ruler with specific objectivity when the model fits. HTA could have adopted these foundations at any time; instead, fixation on QALYs and the valuation of multiattribute health-state descriptions, contrary to the requirement of unidimensionality, guaranteed measurement failure that persists to this day. The remedy is simple and non-negotiable: in HTA there are only two valid measures: linear ratio scales for manifest resource and utilization claims, and Rasch logit ratio scales for latent trait possession.

MODULE 1: THE DENIAL OF FALSIFICATION IN HEALTH TECHNOLOGY ASSESSMENT

Falsification demarcates science by requiring that claims be stated so they can fail against observation. That demands quantities with stable units, so predicted and observed differences are commensurable; explicit conditions under which an expected result would not hold; and the

possibility of replication on the same ruler across settings and time. Representational measurement theory supplies the prerequisite: numbers must preserve an attribute's structure, order, additivity, and invariance, so subtraction, averaging, and ratios are lawful rather than decorative. HTA denies falsification because its cornerstone quantities are not measures. Utilities elicited from time trade-off or standard gamble are ordinal preference indices, yet they are treated as interval or ratio quantities and then multiplied by time to form QALYs. The reference case embeds these non-measures in simulations, tariffs, and thresholds, producing outputs that reflect conventions rather than attributes. Without a validated unit, no observation can disconfirm a claim; models and tariffs change, conclusions endure. That is policy ritual, not science.

MODULE 2: THE RASCH MODEL – LATENT TRAITS AND ITEM SELECTION

This module argues that latent traits, pain, fatigue, mobility, need fulfillment, are scientifically real only when they admit invariant, testable measurement. Representational measurement theory sets the bar: numbers count as measures only when they preserve an attribute's structure under admissible transformations. The Rasch model uniquely delivers this for latent constructs by specifying a single trait, testing items against it, and mapping responses through a logistic function of person location minus item difficulty to place persons and items on a common logit continuum. When data fit, the scale has constant units, preserved order, additivity, solvability, and invariance, enabling lawful arithmetic, hypothesis testing, and falsification. Design follows information: items are most discriminating near a 50% endorsement probability, so instruments target the expected ability region while spanning the continuum to avoid floors and ceilings. Misfit signals instrument or content problems, not a failure of Rasch. In contrast, summed scores and preference utilities remain ordinal encodings that cannot sustain science-ready claims.

MODULE 3: THE RASCH MODEL – THE UNIQUE RASCH LOGIT RATIO SCALE

The creation of a Rasch interval scale is an epistemic requirement, not a statistical convenience. Transforming responses into logits, and logits into an interval ruler, enacts conjecture and refutation: infit, outfit, residual structure, local independence checks, threshold ordering, DIF, and invariance tests probe the axioms of representational measurement. Every misfit is a possible falsification; only by surviving these probes does a latent construct graduate from speculation to measurement. Rasch uniquely operationalizes falsification for latent traits by enforcing order, additivity, and invariance, rejecting instruments that fail. In this sense it is a test of existence: an attribute is measurable only when it yields an invariant scale across persons, items, and time. Specific objectivity, comparisons independent of which well-fitting items or samples are used, marks the point at which numbers earn the name "measure." Following Wright's argument for fundamental measurement, Rasch delivers a tightly coupled dual metric. Additively, logits form a single interval ruler with equal meaning for equal differences; multiplicatively, the same structure yields a ratio metric through odds with a true zero.

MODULE 4: THE RASCH MODEL – POSSESSION AND FALSIFICATION

This module presents possession, the quantitatively expressed amount of a single latent trait, as the primary quantity in Rasch measurement, and the logit as the legitimate scale on which to read it. By modeling ordered responses with Rasch, persons and items are located on a common log-odds continuum; when unidimensionality, ordered categories, local independence, and invariance hold, responses map to an interval ruler where equal differences have equal meaning. Item

difficulty marks required trait; person location marks possessed trait; probabilities follow from their difference. Estimation places persons (θ) and items (β) on this ruler; standard errors indicate precision and enlarge with poor targeting or extreme scores. Precision, coherence, and targeting then determine whether θ merits interpretation as possession. Inference proceeds from person to group: mean change and difference-in-differences are reported on the logit scale, with an odds-ratio translation via $e^{\Delta\theta}$. Linear rescaling aids communication without altering statistics. Anchored calibrations enable before and after claims.

MODULE 5: THE RASCH MODEL - THE EXISTENTIAL CRISIS FOR DISEASE SPECIFIC INSTRUMENTS

Set aside the reference case. The central failure in HTA is more basic: there no patient-reported outcome instrument that meets Rasch measurement requirements. Across disease areas, PROs are universally built from summed ordinal scores of subjective responses and then treated as if they were interval measures. They are not. They routinely lack demonstrated unidimensionality, ordered thresholds, local independence, and sample-free invariance; minimum conditions for a ruler that licenses arithmetic. Without a lawful scale, every subtraction, average, effect size, or regression coefficient built on these totals is numerically incoherent. The field has normalized adding apples to oranges and calling it science.

This is not a technical quibble; it is an indictment. Thousands of HTA practitioners, reviewers, and guideline authors proceed as if numbers were measures by default, ignoring the need to earn additivity through calibration. Checklists, "validations," and psychometric rituals cannot substitute for Rasch construction that conjointly estimates item difficulty and person ability on a common logit scale. Until PRO instruments are Rasch-built and reported as person measures with known error on an invariant ruler, HTA cannot claim to evaluate patient-centered outcomes scientifically. What passes for evidence is, at best, descriptive scoring, incapable of supporting lawful comparisons, change claims, or value assertions. If HTA aspires to be science, its first obligation is clear: replace summed scores with calibrated measures or withdraw patient-reported claims from decision making.

The indictment of measurement in HTA extends beyond the misnamed multi-attribute utility indices (e.g., EQ-5D-3L) to the vast array of disease-specific instruments built on summed scores. From the standpoint of representational measurement theory, these are not measures and cannot lawfully support arithmetic; accordingly, HTA's current corpus of subjective claims is bankrupt. What is needed is not rehabilitation but replacement: Rasch-validated instruments that satisfy unidimensionality, ordered thresholds, local independence, and invariance, yielding person measures on a common logit scale.

HTA has no defensible patient-reported outcome measures. Across disease areas, instruments built from summed ordinal responses are treated as if they were measures, yet they fail the non-negotiable Rasch requirements that would license arithmetic. Without demonstrated unidimensionality, ordered thresholds, local independence, and invariance, a questionnaire yields only response counts on an arbitrary ruler. Numbers are paraded as "scores," then averaged, subtracted, and modeled as though they possessed equal intervals and stable units. They do not. The result is a literature that cannot support evaluable value claims for subjective outcomes because it lacks lawful scales on which change can be located and replicated. This is not a technical

quibble but a categorical failure: without Rasch-validated instruments that place persons and items on a common logit ruler, HTA cannot claim to measure latent constructs at all. The remedy is likewise categorical. Either retire non-measures from decision making, or rebuild the enterprise on Rasch instruments that satisfy conjoint simultaneous measurement and deliver invariant, interval-level person measures. Until then, HTA remains a practice that confuses numbers with measures and forfeits the right to arithmetic, with no basis in science.

CONCLUSION

The conclusion is unavoidable: what has passed for evaluation in health technology assessment is a half-century of numerical storytelling sustained by institutional habit and the seduction of calculation. Numbers were mistaken for measures, simulations for observations, and internal coherence for empirical warrant. The result is a canon of cost-per-QALY ratios and preference utilities that cannot survive the most elementary scrutiny of scale type, additivity, or invariance. When arithmetic is performed on non-measures, precision becomes theater. That is the bankruptcy this program exposes, not as a rhetorical flourish but as a methodological diagnosis grounded in representational measurement theory and the Rasch framework for lawful latent measurement.

The remedy is as clear as it is demanding. First, commit to rulers before results: ratio scales for manifest resource and utilization claims; Rasch-calibrated, invariant logit rulers for latent traits. Second, insist on falsifiable protocols that state, in advance, what would count as failure on the same ruler, in the same fixed target population, within a defined timeframe. Third, prohibit composites and utilities that bundle heterogeneous attributes or reify ordinal preferences; they do not measure anything and cannot lawfully support subtraction, averages, or ratios. Finally, replace model-based narratives with transparent reporting of measured outcomes and their uncertainty. Only then do claims become empirical propositions rather than artifacts of convention.

This transition is not optional for institutions that wish to retain credibility. Formulary committees can either continue to defend a reference-case orthodoxy whose outputs cannot be audited against measurement standards, or they can rebuild evaluation on rulers that earn the right to arithmetic. The former preserves process; the latter restores science. Manufacturers, for their part, can choose to submit dossiers padded with unevaluable scores and projections, or they can design products around lawful endpoints, Rasch instruments where needed, and protocols that permit decisive testing in the real world. Health systems should reward only the second path.

Maimon Research's distance education programs exist to accelerate this reset. They catalog the failures that brought HTA to its present impasse and, more importantly, provide a practical blueprint for measurement-led assessment. Adopting these standards does not constrain inquiry; it liberates it from illusion. Once rulers are fixed and lawful, evidence can accumulate, disagreement can be resolved by observation, and policy can rest on claims that are true or false in the world, not merely reproducible in a model. That is the future

ACKNOWLEDGEMENT

Portions of this working paper were drafted and edited with assistance from ChatGPT (version 5; OpenAI). The author reviewed, verified, and refined all AI-assisted text and assumes full responsibility for the accuracy, integrity, and originality of the final content.

REFERENCES

[1]Stevens S. On the Theory of Scales of Measurement. ;*Science* 1946;103(2684) 677–680

[2] Krantz D, Luce R, Suppes P, Tversky A *Foundations of Measurement, Volume I: Additive and Polynomial Representations.* New York: Academic Press, 1971.

[3] Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Expanded ed. Chicago: University of Chicago Press, 1980 [First printed 1960]

[4] Wright B. Solving Measurement Problems with the Rasch Model. *J Educational Measurement* 1977; 4(2): 97–116

[5] Suppes P. A Set of Independent Axioms for Extensive Quantities. *Portugaliae Mathematica.* 1951; 10: 163–172.

[6] Luce R, Tukey J. Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement *J Mathematical Psychology.* 1964*;* 1(1): 1–27