MAIMON WORKING PAPER No. 16 SEPTEMBER 2025

MEASUREMENT AND FORMULARY SUBMISSIONS: QUESTIONS A COMMITTEE SHOULD ASK

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

ABSTRACT

Formulary committees face the difficult task of evaluating evidence for the value of new therapies. For four decades, however, committees have been presented with claims built on non-measures: utilities, quality-adjusted life years (QALYs), reference case simulations, and patient-reported outcome (PRO)instruments scored by summing ordinal responses. None of these constructs satisfies the axioms of representational measurement theory (RMT). They are not measures. Without interval or ratio properties, numbers cannot support arithmetic operations or statistical inference. Yet the machinery of health technology assessment (HTA) has entrenched these practices, embedding pseudo-quantities at the center of decision-making.

This paper provides a framework for restoring science to formulary review. It begins with an exposition of RMT, explaining in plain terms the axioms of measurement and the representational and uniqueness theorems. It then introduces Rasch measurement as the necessary and sufficient condition for transforming ordinal responses into interval ratio measures, with particular relevance for PROs. The paper reviews the fatal errors of HTA, including utilities, QALYs, and the reference case, and offers a demolition of COSMIN guidelines, which codify the misuse of ordinal data. Against this background, the paper sets out a comprehensive series of questions that formulary committees must ask of submissions, designed to separate credible, evaluable, and replicable claims from pseudo-science.

The conclusion is stark: only two types of measures are admissible in formulary submissions. Manifest constructs must be assessed with linear ratio measures; latent constructs require Rasch logit ratio measures. Everything else utilities, QALYs, composite cost claims, COSMIN checklists, must be rejected. Unless committees enforce these standards, formulary review will remain a conduit for numerical storytelling, complicit in wasting resources and perpetuating error.

Keywords: Formulary committees; Health technology assessment (HTA); Representational measurement theory (RMT); Utilities; Quality-adjusted life years (QALYs); Patient-reported outcomes (PROs); Rasch measurement; COSMIN; Unidimensionality; Ordinal data; Axioms of measurement

INTRODUCTION

Formulary committees are the gatekeepers of health system evidence. Their task is formidable: to weigh the claims made for new therapies, to judge their comparative value, and to decide whether scarce resources should be allocated to them. These are decisions with immense consequences for patients, payers, and society at large. Their legitimacy depends on the quality of the evidence presented.

The problem is that the evidence base has, for decades, failed to meet the most basic requirement of science: measurement. Health technology assessment (HTA), the discipline that has guided formulary submissions, has never embraced the standards of representational measurement theory (RMT) [1] . Instead, HTA has elevated non-measures, utilities, QALYs, reference case models, probabilistic sensitivity analysis and ordinal disease specific patient reported outcomes (PROs) summed scores into received wisdom [2]. These constructs are built on ordinal data and fail every relevant axiom of RMT. Yet committees are asked to treat them as if they were interval or ratio measures.

The failure is not technical but categorical. Numbers that do not preserve empirical relations under admissible transformations are not measures. Arithmetic performed on such numbers is illegitimate. Yet utilities from time trade-off tasks are averaged, multiplied by time, discounted, and fed into simulation models. PROs based on Likert items are summed, correlated, and regressed. These practices have the appearance of quantification but none of it is warranted

This paper provides committees with the tools to resist. It begins with an accessible exposition of RMT, setting out the axioms that define when numbers function as measures. It then introduces Rasch modeling, which provides the necessary and sufficient solution for latent constructs. It reviews the fatal mistakes of HTA, particularly the invention of the QALY, and the further institutionalization of error in COSMIN guidelines. It then sets out a series of questions that committees must ask of every submission. The aim is to ensure that formulary decisions rest on credible, evaluable, and replicable claims; not on numerical storytelling.

REPRESENTATIONAL MEASUREMENT THEORY: FOUNDATIONS AND AXIOMS

Measurement is the foundation of science. It is what allows us to move from description to quantification, from impression to evidence. But not all numbers are measures. RMT formalized by 1971 makes clear this distinction. At its core, RMT asks: under what conditions can empirical relations be faithfully represented numerically? To answer this, RMT identifies a set of axioms or rules that the empirical system must satisfy. When these axioms hold, numbers can preserve the structure of relations, and measurement is possible. When they do not, numbers are merely labels or ranks, incapable of supporting arithmetic or inference.

The axioms of measurement are the bedrock of representational measurement theory. The first is the axiom of order. This requires that if one object is empirically greater than another, the number assigned to it must also be greater. If patient A survives longer than patient B, then the numerical value assigned to A must exceed that assigned to B. Without order, numbers lose their ability to represent even the simplest comparative relations. The axiom of additivity goes further. If combining two quantities empirically produces a third, then the corresponding numbers must combine by arithmetic addition. Two tablets of 100 mg each must equal one of 200 mg. Without additivity, the arithmetic structure breaks down, and numbers cannot reflect empirical composition. The axiom of solvability requires that for any two magnitudes there must exist an intermediate magnitude. If one patient survives 12 months and another 24, there must be the possibility of a patient surviving 18 months. Solvability is what makes continuity possible and allows us to interpolate between observations. The cancellation axioms ensure that equivalence is respected. If the combination of hospital stays A and B is empirically the same as the combination

of hospital stays C and D, then the numbers assigned must balance in the same way. This property ensures that empirical equivalences are preserved arithmetically. The Archimedean property provides the final safeguard of proportionality. It requires that small units, added enough times, must eventually exceed larger ones. A day, added enough times, must eventually surpass a week. Without this property, the scale would collapse into non-comparability.

These axioms are not conventions to be adopted or ignored as convenience dictates. They define the very conditions under which numbers legitimately function as measures. If the axioms are not met, then numbers are nothing more than labels or ranks.

When these axioms are satisfied, the representational theorem guarantees that there exists a numerical mapping that preserves the observed empirical relations. This theorem does not merely state that measurement is desirable; it proves that if the axioms are met, then a scale exists that represents the structure of the empirical system in numbers. The importance of this theorem lies in its generality: it shows that measurement is possible whenever empirical systems conform to the axioms, and it gives science its quantitative backbone.

Equally important is the uniqueness theorem. This theorem identifies what transformations of numbers preserve meaning once measurement has been established. For interval scales, meaning is preserved under linear transformations. One can change the zero point or alter the size of the unit, but differences remain valid. This is why temperature can be expressed in Celsius or Fahrenheit without altering its interval structure. For ratio scales, meaning is preserved only under multiplication by a positive constant. Doubling or halving the scale does not alter ratios, but shifting the zero point is not permissible because the presence of a true zero is fundamental. This is why weight can be measured in kilograms or pounds, but not in a scale that arbitrarily sets zero at a non-zero weight.

Together, the representational and uniqueness theorems establish the boundary between measurement and mere numerical convenience. RMT guarantees that measurement is possible if the axioms hold. The uniqueness theorem specifies the transformations that maintain validity once measurement is established. In tandem, they define not only when measurement exists but also what operations are legitimate on the resulting numbers. For applied fields such as health technology assessment and formulary review, these theorems have direct implications. Arithmetic operations on data are legitimate only when the axioms are satisfied and the transformations identified by the uniqueness theorem are respected. If these conditions are not met, then numbers cannot be treated as measures, and any claims based on them collapse into pseudo-science.

In 1946, Stevens published what has become one of the most cited papers in the social sciences: *On the Theory of Scales of Measurement* [3] His aim was pragmatic. Scientists were already using numbers in many ways, but there was little clarity about what different kinds of numbers meant, or what statistical operations were legitimate on them. Stevens proposed a typology of four scale types, nominal, ordinal, interval, and ratio, that provided a simple way to classify numerical assignments according to the properties they preserved. This typology has been enormously influential, not least because it seemed to offer a ready-made justification for the expanding use of statistical methods in psychology and the social sciences.

Nominal scales were defined as pure labels. They allowed classification but no ordering. Ordinal scales preserved order, but not the equality of differences. Interval scales went further, preserving equality of differences but not ratios, since they lacked a true zero. Ratio scales preserved both equal intervals and ratios, adding the fundamental property of a true zero. Stevens's genius was to link this classification to the legitimacy of statistical operations. With nominal scales, only counts and modes were appropriate. With ordinal scales, one could rank, but not add or average meaningfully. Interval scales permitted the use of means, variances, and correlation coefficients. Ratio scales admitted the full range of arithmetic, including multiplication and division.

This framework was pivotal. It drew attention to the fact that different kinds of scales allow different kinds of inferences, and that not all numbers are created equal. More importantly, it gave researchers a working set of categories to justify the use of statistics. If an instrument could plausibly be treated as interval, then the door was open to apply the full battery of statistical methods. In this way, Stevens's typology provided the bridge between practical measurement and statistical analysis.

But Stevens's scheme was descriptive and pragmatic, not axiomatic. It offered categories but not the logical structure that made those categories rigorous. What it lacked was a demonstration of when empirical relations genuinely admit interval or ratio representations. This is where representational measurement theory entered. In the decades after Stevens, mathematicians and psychologists including Luce, Tukey, Krantz, Suppes, and Tversky developed the axiomatic foundations that could justify Stevens's categories [4] [5]. By identifying the axioms of order, additivity, solvability, cancellation, and the Archimedean property, they showed precisely when a numerical representation exists and what transformations preserve its meaning. The representational theorem established the conditions for measurement; the uniqueness theorem identified the admissible transformations.

In effect, RMT supplied the theoretical backbone that Stevens's typology had anticipated but could not provide. Stevens had pointed to the importance of interval and ratio scales for statistical inference. RMT explained why: only when the axioms are satisfied do numbers qualify as interval or ratio measures, and only then are arithmetic operations and statistical analyses legitimate. The pivotal role of Stevens's typology was to provide the intellectual bridge between practical measurement tasks and the later formalism of RMT. His classification made clear that interval and ratio measures were essential for science, while RMT transformed this insight into a set of rigorous theorems that grounded measurement in logic and mathematics.

For formulary committees, the continuity from Stevens to RMT matters. The statistical analyses routinely applied to utilities, QALYs, and PRO scores incorrectly presume interval or ratio properties. Stevens's typology already made clear that such operations are illegitimate on ordinal data. RMT sharpened the point: unless the axioms are satisfied, no amount of statistical manipulation can transform an ordinal index into a measure. What Stevens provided in pragmatic form, RMT confirmed in axiomatic rigor. Together, they draw the line between science and pseudo-science.

RASCH MEASUREMENT: THE SOLUTION FOR LATENT CONSTRUCTS

Manifest constructs such as survival time or blood pressure yield ratio measures directly. But many constructs of interest in health care are latent: fatigue, pain, satisfaction, need fulfillment. These are not directly observable. They manifest through responses to items. Typically, these responses are ordinal—"none," "some," "severe"—and are scored using Likert categories. Summing these responses into totals does not transform them into measures. They remain ordinal.

Georg Rasch proposed in the 1950s a probabilistic model for responses [6]. The probability of a given response depends on the difference between the person's ability (or trait level) and the item's difficulty (or severity). This probability is expressed as a logistic function. [i.e. the natural logarithm of the odds ratio ln(p/1-p]. Applying this across respondents and items yields a unidimensional continuum, expressed in logits. Both persons and items are located on the same scale.

The Rasch model ensures that a latent trait is measured along a single continuum, with differences on the logit scale representing constant proportional changes regardless of position. It produces invariant results: person estimates are independent of the specific items used, and item calibrations are independent of the particular sample of persons. By locating both persons and items on the same scale, Rasch achieves conjoint simultaneous measurement, a property that Wright demonstrated in 1977 to be equivalent to the axioms of representational measurement theory [7]. Rasch is therefore not a statistical convenience or approximation but a realization of measurement itself [8].

It is important to note that with Rasch measurement the impact of therapy interventions is captured by the individual or group possession of a manifested latent trait. This is defined in logits in terms of the Rasch logit number line. This is the only statistic that represents for PROs the impact of therapy interventions. Changes in the manifested latent trait are evaluated with established statistical techniques in both interval and ratio forms. Unlike summed scores, which assume but never demonstrate interval properties, Rasch calibrations create a unidimensional continuum where both item difficulty and person ability are placed on the same invariant scale. Possession of the trait is quantified, not assumed, and movement along the logit scale has constant meaning across the measurement range. This enables legitimate arithmetic and statistical inference. In the context of therapy evaluation, Rasch measures provide the only defensible way to track true change in patient outcomes.

This stands in sharp contrast to traditional psychometric methods such as Cronbach's alpha, factor analysis, or correlation studies. These procedures merely describe associations within ordinal data and assume interval properties that are never demonstrated. Rasch, by transforming ordinal responses into interval logit measures, uniquely satisfies the requirements of measurement theory.

For formulary committees, the implications are decisive. In the case of latent constructs, Rasch modeling is the necessary and sufficient condition for scientific adequacy. Unless PROs have been calibrated using Rasch, their scores remain ordinal totals that cannot legitimately support arithmetic or inference. Committees must therefore ask whether Rasch modeling has been applied, and if the answer is no, the claim should be rejected.

THE FATAL MISTAKES OF HTA: UTILITIES, QALYS AND THE REFERENCE CASE

By the late 1970s the foundations of representational measurement theory were well established. The axioms had been formalized, the representational and uniqueness theorems articulated, and it was already clear that multiattribute indices could never yield interval or ratio measures. The reason is straightforward but decisive: measurement requires unidimensionality. A scale can only represent one underlying construct, preserving order, additivity, solvability, cancellation, and invariance along a single continuum. Once multiple attributes are bundled together, such as mobility, pain, self-care, anxiety, and usual activities, the requirements of measurement collapse. No multiattribute construct can ever yield a valid scale because it fails the most basic condition of unidimensionality. There is, quite simply, no such thing as a multiattribute measure.

The so-called "preferences" generated by health state valuation exercises are not ordinal in any meaningful sense; they are multi-attribute constructs with no foundation in measurement theory. Despite this, HTA proceeded as if they could be elevated to the status of measures. Under pressure to produce a single number to guide resource allocation, it embraced time trade-off and standard gamble tasks applied to descriptive profiles of health states. These were never instruments of measurement but devices for eliciting hypothetical judgments. The results were then forced through regression algorithms to produce scoring systems such as the EQ-5D-3L. Yet the fundamental defect was never addressed. The outputs remained nothing more than multi-attribute indices masquerading as measures, incapable of meeting the axioms of representational measurement theory. They lacked interval equality, true zeros, invariance, or any property required for credible scientific claims, and their continued use represents a profound methodological failure.

Yet these indices were christened "utilities" and given a spurious legitimacy by discounting time in a disease state to give quality-adjusted life years. This was a category mistake of the first order. Survival time is a ratio measure: it is unidimensional, admits units, possesses a true zero, and supports ratio comparisons. Utilities derived from multiattribute health state descriptions are not meaningful indices; they are not even ordinal. Multiplying a ratio measure by an index violates the principle of dimensional homogeneity. The resulting QALY is not a measure at all but a pseudo-quantity, a number with the appearance of arithmetic but no equivalence in measurement terms.

The final culmination of these errors was the reference case model, advanced as the gold standard for cost-effectiveness analysis. Built upon QALYs, it inherited their categorical flaws. The outputs of reference case models cannot be replicated or falsified because they are generated from inputs that are not measures. The models are not designed to test claims but to produce numbers, giving the illusion of quantification while institutionalizing numerical storytelling.

For formulary committees the lesson is unavoidable. The pursuit of multiattribute constructs guaranteed failure because it sought to create measures where none could exist. By embedding utilities, QALYs, and reference case models at the heart of submissions, HTA condemned itself to decades of pseudo-science. No matter how apparently comprehensive the methodology or how widely accepted the practice, these constructs cannot be rescued. They are scientifically indefensible. Submissions based on them must be recognized for what they are, non-measures dressed as measures, and rejected.

COSMIN AND THE APOTHEOSIS OF THE ORDINAL SCALE

The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative is widely promoted as the international benchmark for evaluating patient-reported outcome measures [9]. Its guidelines are often cited in formulary submissions as proof that an instrument is valid and reliable. Yet COSMIN represents not a safeguard but the apotheosis of measurement failure. It elevates the misuse of ordinal data into a formal system, presenting statistical rituals as if they were scientific guarantees. In doing so, it enshrines nonsense at the center of methodological orthodoxy. One might say COSMIN is measurement nonsense on stilts, raising error to new heights while demanding that committees bow to its authority.

The COSMIN framework sets out a checklist of "measurement properties" such as content validity, reliability, structural validity, responsiveness that sound authoritative. But each is assessed through statistical tests on summed ordinal scores. Cronbach's alpha, factor analysis, correlations, and hypothesis testing are deployed as evidence of quality, even though each presupposes interval properties that PRO data lack. Stevens made this clear in 1946. There is no requirement to demonstrate unidimensionality in the Rasch sense, no recognition of invariance, no concern for solvability, cancellation, or additivity. COSMIN bypasses the axioms of RMT entirely.

The result is parallel to the QALY fiasco. Just as utilities from time trade-off tasks are ordinal indices masquerading as interval values, COSMIN validates summed PRO scores as if they were measures. Both frameworks mistake statistical association for quantification and institutionalize pseudo-science. The difference is that COSMIN codifies the error into international guidance, elevating the ordinal scale, the weakest of scales, to a status it cannot sustain. After some 80 years since Stevens and over 50 years since the formalization of RMT, it is puzzling that the penny had not dropped on the limitations of ordinal scales; but COSMIN is not alone.

For formulary committees, the warning could not be clearer. A claim justified by reference to COSMIN is not strengthened but undermined. COSMIN validation does not and cannot transform ordinal totals into interval or ratio measures. Unless a PRO has been calibrated with Rasch modeling, its scores are not measures, no matter how many COSMIN boxes have been ticked. Committees should treat citation of COSMIN not as a mark of rigor but as a red flag, signaling that the submission is rooted in non-measures and must be rejected as scientifically indefensible.

QUESTIONS A FORMULARY COMMITTEE SHOULD ASK

The theory is clear. But how should committees apply it? The following questions translate measurement standards into practice. Each question is a filter designed to expose pseudo-claims.

- **What is being measured?**

The first step is classification. Is the construct manifest or latent? If manifest (e.g., survival, blood pressure, tumor size), ratio measures must be presented. If latent (e.g., pain, fatigue), Rasch logit measures are required. Submissions must make this distinction explicit.

- **Do the data satisfy the axioms of RMT?**

Submissions must be interrogated against the axioms of RMT. Is the construct unidimensional? Are values additive and invariant? Are cancellation and solvability demonstrated? If not, the numbers are not measures. Committees must demand evidence.

- **What scale type is claimed?**

If interval or ratio is claimed, proof must be presented. Many instruments assert interval properties without evidence. PROMs, utilities, and composite scores fall into this trap. Committees must require demonstration, not assertion. Is the scale linear ratio of Rasch ratio logit?

- **Has Rasch modeling been applied?**

For latent construct traits, Rasch is the only admissible pathway. Has the instrument been Rasch-calibrated? Are item and person parameters invariant? Is unidimensionality demonstrated? Will possession of the latent trait be demonstrated? If not, the submission must be rejected.

- **Are value claims credible, evaluable, replicable?**

Measurement is not an end in itself. Claims must be testable. Can the claims be evaluated in practice? Could it be replicated by others? Utilities, QALYs, and summed scores fail this test. There are only two acceptable measures for claims: linear ratio claims with relative absolute difference and Rasch logit ratio claims with constant relative differences.ro

- **Does the submission rely on utilities, QALYs, reference cases?**

If yes, this is a red flag. Utilities are ordinal; QALYs violate dimensional homogeneity; reference cases are not falsifiable. Claims based on them are invalid.

- **How are PROMs presented?**

PROMs are common in submissions. Are they Rasch-calibrated, producing logit ratio claims? Or are they summed scores, producing ordinal totals? Only the former are admissible.

- **Are resource utilization claims ratio-based?**

Claims should be expressed in as linear ratio measures: hospital days, ER visits, readmissions. Composite “cost” claims that bundle heterogeneous resources are invalid.

- **What timeframe is specified?**

Credible claims must be anchored in evaluable timeframes. Committees should ask: can this claim be tested within 12 months of launch? If not, it risks being speculative.

- **Are comparators defined and protocols specified?**

For replication and evaluation, comparator products must be identified and protocols agreed. Without these, claims lack context.

CONCLUSIONS

After forty years of failure, there can be no more equivocation: the axioms of representational measurement theory are non-negotiable. They are not academic curiosities or optional guidelines; they are the necessary and sufficient conditions for turning numbers into measures. To continue ignoring them is to perpetuate a scientific fraud. Health technology assessment has been built on a foundation of meaningless scores masquerading as interval data, utilities constructed from time trade-off preferences, and the grotesque chimera of the QALY. These are not lapses that can be corrected at the margin. They are category mistakes, fatal from inception, and they have produced four decades of numerical storytelling under the false banner of science.

The lesson is clear: formulary value claim submissions cannot admit any instrument, any construct, or any model that fails the axioms. Claims must be credible, evaluable, and replicable. Credibility is impossible without unidimensionality, interval scaling, invariance, and a true zero where required. Evaluability is impossible if claims are built on indices that dissolve under the most elementary cancellation tests. Replicability is impossible if results are driven by algorithms, tariffs, and mapping tricks that have no measurement warrant. The very integrity of science demands that these requirements be enforced.

The EQ-5D is the clearest example of failure: a multi-attribute ordinal index promoted as a utility, then multiplied by time to form the QALY, a construct that is mathematically indefensible. Likewise, the COSMIN framework has elevated summed ordinal PROM scores to the apotheosis of measurement nonsense. These are not “instruments” but artifacts of statistical wishful thinking, incapable of yielding interval measures or meeting the most basic axioms of RMT. They have no place in formulary submissions, no matter how entrenched they have become.

The defense that “this is how HTA has always been done” is not an argument but an indictment. Forty years of entrenched practice has not created legitimacy; it has compounded error. Health systems, patients, and manufacturers deserve better than the pretense of measurement. They deserve standards consistent with normal science. The only viable way forward is a total reset: the immediate abandonment of meaningless scores and the adoption of protocols that honor the axioms of RMT.

The time for compromise has passed. A formulary submission that ignores measurement axioms is not incomplete; it is invalid. The axioms of representational measurement are not negotiable, and value claims that fail them are dead on arrival. HTA must confront its history of error, admit the collapse of its core constructs, and set new standards grounded in science. Anything less is capitulation to pseudoscience. The future of evidence-based decision making depends on the courage of formulary committees to draw this line.

ACKNOWLEDGEMENT

Portions of this education program 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] Krantz D, Luce R, Suppes P, Tversky A**.** *Foundations of Measurement*, Volumes I–III. New York: Academic Press, 1971 (Vol. I); 1989 (Vol. II); 1990 (Vol. III).

[2] Drummond M, Sculpher M, Claxton K et al. Methods for the Economic Evaluation of Health Care Programmes (4th Ed.) New York: Oxford University Press, 2015

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

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

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

[6] Rasch, Georg. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded ed., Chicago: University of Chicago Press, 1980)

[7] Wright B. "Solving Measurement Problems with the Rasch Model." *Journal of Educational Measurement* 14, no. 2 (1977): 97–116

[8] Bond T, Zi Yan, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences (4th Ed). New York: Routledge, 2021

[9] Mokkink L, Elsman E, Terwee C. COSMIN Guideline for Systematic Reviews of Patient-Reported Outcome Measures Version 2.0. *Quality Life Research*. 2024; 33(11); 2929–39