MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

THE MEASUREMENT FUTURE IN HEALTH TECHNOLGY ASSESSMENT

NORWAY: MEASUREMENT FAILURE IN LEGACY HEALTH TECHNOLOGY ASSESSMENT

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

LOGIT WORKING PAPER No 1078 AUGUST 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

Background: *Norway has developed a highly organized national health technology assessment (HTA) environment through the New Methods (Nye metoder) system, integrating assessment, appraisal, economic evaluation and priority setting. This procedural sophistication provides an important setting for examining whether the quantitative methods of legacy HTA satisfy the requirements of representational measurement theory (RMT).*

Objectives: *To assess whether the Norwegian national HTA knowledge base possesses the measurement principles required to justify its established arithmetic and whether its treatment of utilities, quality-adjusted life years (QALYs), cost-effectiveness, latent attributes and reference-case simulations is consistent with RMT.*

Methods: *A large language model interrogation applied 24 canonical TRUE/FALSE statements derived from RMT and Rasch measurement principles to the Norwegian HTA knowledge environment. Categorical endorsement probabilities estimated the extent to which each proposition is reinforced explicitly or implicitly within the knowledge base. Probabilities were transformed to normalized logits ranging from −2.50 to +2.50. The unit of analysis was the cumulative Norwegian HTA knowledge environment rather than individual practitioners.*

Results: *The findings demonstrate pronounced measurement inversion. The TRUE propositions that measurement precedes arithmetic, multiplication requires ratio measurement and RMT axioms are required for arithmetic each received probabilities of 0.10 (−2.20 logits). In contrast, the FALSE proposition that the QALY is a ratio measure received 0.95 (+2.50), while FALSE propositions that the QALY is dimensionally homogeneous and can be aggregated received 0.90 (+2.20) and 0.95 (+2.50), respectively. Time was strongly recognized as a ratio measure, 0.95 (+2.50), despite weak recognition of the ratio requirement governing its multiplication by utility. Rasch measurement principles for latent attributes were essentially absent, with key items at 0.05 (−2.50). Reference-case simulations were strongly treated as generating falsifiable claims, 0.90 (+2.20), despite their conditional model structure.*

Conclusions: *Norway demonstrates that procedural discipline cannot substitute for measurement discipline. A highly managed national HTA process has institutionalized QALY and cost-effectiveness arithmetic without comparably institutionalizing the measurement requirements necessary to establish its scientific admissibility. The required transition is not toward greater modelling sophistication but toward restoration of the scientific sequence: attribute, measurement, scale, admissible arithmetic, therapy-impact claim, empirical evaluation, replication and falsification. Measurement must become the gatekeeper of Norwegian quantitative HTA.*

INTRODUCTION

Norway provides an unusually revealing setting in which to examine the measurement foundations of legacy health technology assessment (HTA). Its national HTA environment is characterized by a high degree of procedural organization, centralized coordination and explicit priority-setting. The New Methods (*Nye metoder*) system was established to provide a structured national process through which new health technologies entering the specialist health service are assessed, appraised and considered for public funding. Benefit, resource use and severity provide explicit criteria for priority setting, while evidence synthesis, economic evaluation, modelling and uncertainty analysis contribute to a highly organized assessment process.

This institutional sophistication makes the Norwegian case particularly important. The scientific question is not whether Norway possesses an orderly HTA process. It clearly does. Nor is the question whether its assessments employ sophisticated quantitative methods. They do. The prior question is whether the quantities entering those methods possess the measurement properties required for the arithmetic performed upon them [1] [2]

Representational measurement requires arithmetic to be governed by the properties of the measurement scale. Numerical values cannot legitimately be multiplied, divided, added or aggregated merely because numbers have been assigned to observations. Measurement must first be established, the relevant scale properties demonstrated and the proposed arithmetic shown to be admissible. The distinction between nominal, ordinal, interval and ratio scales is therefore not a matter of terminology. It determines which mathematical operations possess empirical meaning.

This requirement has immediate consequences for Norwegian HTA. Cost-utility analysis occupies an important position within the national framework. Preference-weighted health-state values are combined with time to generate quality-adjusted life years (QALYs); differences in QALYs are calculated; costs are aggregated; and incremental costs are divided by incremental QALYs to produce cost-per-QALY ratios. Decision models subsequently combine these and other numerical inputs through increasingly elaborate sequences of arithmetic.

Each operation presupposes measurement. The familiar expression $QALY = U \times T$, for example, requires multiplication. Time can satisfy the requirements of linear ratio measurement. The critical question is therefore whether the utility value possesses the measurement properties necessary to operate as a dimensionless proportional adjustment to time. Those properties cannot be conferred upon utility by the fact that Norwegian HTA requires the multiplication. They must be demonstrated independently before the multiplication is undertaken.

The same principle applies downstream. If utility has not been established as an admissible quantitative measure, multiplication by time cannot manufacture a QALY measure. Subtracting QALYs to generate an incremental effect does not repair the initial failure, and dividing incremental costs by an inadmissible ΔQALY cannot create a scientifically meaningful cost-per-QALY ratio. Sensitivity analysis, scenario analysis, probabilistic modelling and increasingly sophisticated simulation cannot manufacture measurement properties absent from the quantities entering the model.

Latent attributes create a related measurement requirement. Pain, fatigue, physical function, need fulfilment and similar attributes are not directly observable manifest quantities. Assigning numbers to questionnaire responses, summing response categories or applying preference weights does not by itself establish measurement. A measurement model must demonstrate a defined unidimensional attribute, invariance and the required quantitative structure. Within the framework applied here, Rasch measurement provides the pathway for constructing an invariant logit continuum for a latent attribute and establishing possession of that attribute.

The Norwegian interrogation therefore asks a question that precedes the procedural architecture of *Nye metoder*: does the national HTA knowledge base possess the measurement principles necessary to determine whether its established arithmetic is scientifically admissible?

The results reveal a pronounced asymmetry. Principles that should govern arithmetic including measurement preceding arithmetic, the ratio requirement for multiplication and the axioms of representational measurement are weakly represented. In contrast, propositions necessary to sustain conventional QALY arithmetic are strongly reinforced even where those propositions are false under representational measurement. The Norwegian knowledge base strongly recognizes time as a ratio measure while scarcely recognizing the ratio-measurement requirement governing the multiplication of time by utility.

This is measurement inversion. The normal scientific sequence attribute, measurement, scale and then admissible arithmetic has been reversed. Arithmetic has become an established component of a highly disciplined national process without a prior demonstration that the numerical constructs entering that arithmetic possess the required measurement properties.

Norway therefore presents a particularly important paradox. It has created a managed national process for the introduction of health technologies while leaving the measurement foundations of its central quantitative constructs effectively unmanaged. Procedural discipline cannot substitute for measurement discipline. The purpose of this assessment is to determine how deeply that inversion is embedded within the Norwegian national HTA knowledge base and what the findings imply for the future of quantitative therapy evaluation in Norway.

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 NORWEGIAN NATIONAL HTA KNOWLEDGE BASE

The Norwegian national HTA knowledge base is best understood as an integrated institutional and methodological environment rather than the output of any single organization. It encompasses the national assessment and priority-setting framework, the *Nye metoder* process, health-economic evaluation, methodological guidance, assessment reports, evidence-synthesis procedures, modelling conventions and the wider academic and professional literature that supports technology evaluation in Norway. Collectively, these sources define the quantitative language through which evidence is generated, interpreted and translated into decisions.

A defining feature of this environment is its high degree of organization. The Norwegian system was designed to provide a consistent national pathway for introducing new technologies into the specialist health service. Assessment, appraisal and decision-making are linked within a structured process involving defined institutional responsibilities and the regional health authorities. Priority setting is organized around benefit, resource use and severity, giving the system an explicit framework within which competing claims on healthcare resources can be considered.

Economic evaluation is embedded within this structure. Cost-utility analysis provides a major quantitative mechanism for relating health outcomes to resource consequences. Preference-based health-state values, commonly generated through multiattribute instruments such as EQ-5D, are combined with time to produce QALYs. Incremental costs and incremental QALYs are subsequently combined to generate cost-per-QALY ratios. Models may extrapolate outcomes beyond observed data, incorporate assumptions about survival and health-state values and explore uncertainty through deterministic, probabilistic and scenario analyses.

The resulting knowledge base is therefore quantitatively sophisticated. Assumptions are documented, alternative scenarios can be examined, uncertainty can be characterized and model structures can be subjected to extensive sensitivity analysis. These characteristics demonstrate considerable procedural discipline. They do not, however, establish the measurement status of the quantities entering the analysis.

That distinction defines the present interrogation. The relevant question is not whether Norwegian analysts competently apply the methods prescribed by the established HTA framework. Nor is the interrogation an assessment of the knowledge or competence of individual researchers, analysts or decision-makers. The unit of analysis is the publicly represented Norwegian HTA knowledge environment: the cumulative body of methodological rules, conventions, assumptions and practices through which quantitative HTA is expressed and reproduced.

This distinction is essential. A high endorsement probability for a FALSE canonical statement does not mean that Norwegian practitioners consciously believe a false proposition. It means that the knowledge environment behaves as though that proposition were acceptable through its methods, assumptions or repeated use. Conversely, weak endorsement of a TRUE measurement proposition indicates that the principle is rarely articulated, required or allowed to constrain established quantitative practice.

Against this standard, the central issue is whether representational measurement functions as a gatekeeper within Norwegian HTA. Before utilities are multiplied by time, does the knowledge base require demonstration that utility possesses the necessary measurement properties? Before QALYs are aggregated or differences between QALYs calculated, is their scale status established? Before incremental costs are divided by incremental outcomes, is the denominator demonstrated to support the required arithmetic? For latent attributes, is a measurement model required to establish a defined unidimensional attribute and an invariant quantitative continuum before numerical claims of therapy impact are made?

The Norwegian knowledge base examined here provides little evidence that these questions occupy a logically prior position. Instead, the conventional arithmetic of cost-utility analysis appears as an already established component of the system. Preference-weighted indices are treated as utilities, utilities are multiplied by time, QALYs are aggregated, incremental QALYs are calculated and cost-effectiveness ratios are generated. The measurement properties necessary to justify these operations are largely assumed rather than independently demonstrated.

This creates an important distinction between process control and measurement control. Norway has developed an unusually structured system for controlling how technologies enter assessment, how evidence is reviewed, how economic models are presented and how decisions are reached. Yet representational measurement does not appear to exercise comparable control over what numerical operations may legitimately enter those analyses.

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

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.

ASSESSMENT: MEASUREMENT INVERSION IN NORWEGIAN HTA

The Norwegian probability and normalized-logit profile presents an exceptionally clear example of measurement inversion (Table 1). It does not describe a knowledge base in which representational measurement is substantially understood but occasionally overlooked in practice. Nor does it suggest a marginal disagreement over one or two technical features of economic evaluation. The pattern is systematic. Principles that should determine whether quantitative operations are admissible receive some of the weakest endorsements in the interrogation, while propositions required to sustain the conventional arithmetic of legacy HTA receive some of the strongest. The result is not simply an omission of measurement theory from an otherwise coherent analytical framework. It is a reversal of the scientific sequence through which quantitative claims should acquire empirical meaning

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

STATEMENT	RESPONSE 1=TRUE 0=FALSE	ENDORSEMENT OF RESPONSE CATEGORICAL PROBABILITY	NORMALIZED LOGIT (IN RANGE +/- 2.50)
INTERVAL MEASURES LACK A TRUE ZERO	1	0.20	-1.40
MEASURES MUST BE UNIDIMENSIONAL	1	0.15	-1.75
MULTIPLICATION REQUIRES A RATIO MEASURE	1	0.10	-2.20
TIME TRADE-OFF PREFERENCES ARE UNIDIMENSIONAL	0	0.85	+1.75
RATIO MEASURES CAN HAVE NEGATIVE VALUES	0	0.90	+2.20
EQ-5D-3L PREFERENCE ALGORITHMS CREATE INTERVAL MEASURES	0	0.90	+2.20
THE QALY IS A RATIO MEASURE	0	0.95	+2.50
TIME IS A RATIO MEASURE	1	0.95	+2.50
MEASUREMENT PRECEDES ARITHMETIC	1	0.10	-2.20

SUMMATIONS OF SUBJECTIVE INSTRUMENT RESPONSES ARE RATIO MEASURES	0	0.90	+2.20
MEETING THE AXIOMS OF REPRESENTATIONAL MEASUREMENT IS REQUIRED FOR ARITHMETIC	1	0.10	-2.20
THERE ARE ONLY TWO CLASSES OF MEASUREMENT LINEAR RATIO AND RASCH LOGIT RATIO	1	0.05	-2.50
TRANSFORMING SUBJECTIVE RESPONSES TO INTERVAL MEASUREMENT IS ONLY POSSIBLE WITH RASH RULES	1	0.05	-2.50
SUMMATION OF LIKERT QUESTION SCORES CREATES A RATIO MEASURE	0	0.95	+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	0.60	+0.40
REFERENCE CASE SIMULATIONS GENERATE FALSIFIABLE CLAIMS	0	0.90	+2.20
THE LOGIT IS THE NATURAL LOGARITHM OF THE ODDS-RATIO	1	0.60	+0.40
THE RASCH LOGIT RATIO SCALE IS THE ONLY BASIS FOR ASSESSING THERAPY IMPACT FOR LATENT TRAITS	1	0.05	-2.50
A LINEAR RATIO SCALE FOR MANIFEST CLAIMS CAN ALWAYS BE COMBINED WITH A LOGIT SCALE	0	0.40	-0.45
THE OUTCOME OF INTEREST FOR LATENT TRAITS IS THE POSSESSION OF THAT TRAIT	1	0.25	-1.10
THE RASCH RULES FOR MEASUREMENT ARE IDENTICAL TO THE AXIOMS OF REPRESENTATIONAL MEASUREMENT	1	0.05	-2.50

The proposition that measurement precedes arithmetic is TRUE yet receives an endorsement probability of only 0.10 (−2.20 logits). The proposition that multiplication requires a ratio measure is also TRUE and receives the same 0.10 (−2.20). The requirement that the axioms of representational measurement must be satisfied for arithmetic receives only 0.10 (−2.20), while unidimensionality, another fundamental measurement requirement, receives 0.15 (−1.75). These are not specialist propositions at the margins of quantitative science. They determine whether numerical operations are empirically meaningful. Their weak representation means that the Norwegian HTA knowledge base possesses an elaborate quantitative apparatus without strongly possessing the rules required to determine whether that apparatus is operating upon measures.

The contrast with the QALY is extraordinary. The proposition that the QALY is a ratio measure is FALSE under the representational-measurement framework applied here, yet it receives an endorsement probability of 0.95 (+2.50 logits), the maximum normalized positive value. The proposition that the QALY is dimensionally homogeneous, also FALSE, receives 0.90 (+2.20), while the proposition that QALYs can be aggregated, again FALSE, receives 0.95 (+2.50). The knowledge base therefore exhibits almost the mirror image of the structure that normal quantitative science would require. The propositions capable of questioning or constraining QALY arithmetic are strongly negatively represented, while the propositions required to permit and perpetuate that arithmetic are strongly positively represented. The directional asymmetry is too consistent to be dismissed as a collection of unrelated omissions.

Perhaps the most revealing result is the relationship among the measurement status of time, the requirement for ratio measurement before multiplication and the assumed status of the QALY. Norway strongly possesses the TRUE proposition that time is a ratio measure, with an endorsement probability of 0.95 (+2.50). Yet the equally fundamental TRUE proposition that multiplication requires a ratio measure receives only 0.10 (−2.20), while the FALSE proposition that the QALY is a ratio measure receives 0.95 (+2.50). This contradiction goes directly to the construction of the QALY. In the familiar expression QALY = U × T, Norwegian HTA recognizes the measurement status of one operand, time, while scarcely recognizing the measurement requirement governing the multiplication itself. The fact that time is a ratio measure does not establish that utility is a ratio measure. Nor can multiplication by a ratio measure confer ratio properties upon a second quantity that has not independently been demonstrated to possess them.

The required scientific sequence would be straightforward. The attribute represented by utility should first be defined. Its measurement structure should then be established, its scale properties demonstrated and the admissible transformations determined. Only after establishing that utility possesses the properties necessary to operate as a dimensionless proportional adjustment to time would multiplication be scientifically defensible. Legacy HTA reverses this sequence. The multiplication is accepted first, the numerical product is called a QALY, and the QALY is subsequently treated as though the arithmetic that created it were evidence of its measurement status. It is not. The ability to calculate U × T demonstrates only that two numbers can be multiplied. It does not demonstrate that the resulting number represents an empirically meaningful quantity.

This distinction between numerical assignment and measurement is reinforced elsewhere in the Norwegian profile. The FALSE proposition that time trade-off preferences are unidimensional receives an endorsement probability of 0.85 (+1.75). The FALSE proposition that EQ-5D-3L preference algorithms create interval measures receives 0.90 (+2.20), while the FALSE proposition that ratio measures can have negative values receives 0.90 (+2.20). These results identify a knowledge environment in which the generation of numerical values has effectively displaced the demonstration of measurement. A preference elicitation procedure can generate numbers. A valuation algorithm can assign numbers to health states. A scoring rule can transform responses into numerical indices. None of these operations establishes that the resulting values constitute measures. Measurement status depends upon the empirical and mathematical structure demonstrated for the attribute, not upon the sophistication of the procedure used to generate numbers.

This is particularly important for preference-based multiattribute instruments. A numerical algorithm applied to combinations of health-domain responses does not demonstrate a unidimensional attribute, equal units, invariance or a non-arbitrary zero. Calling the resulting values utilities does not alter their measurement properties. Anchoring selected values at zero and one does not establish a ratio scale. Preference elicitation does not manufacture the structure required for multiplication. Yet the Norwegian HTA framework proceeds operationally as though these prior questions have already been resolved. Repeated numerical assignment has effectively become a substitute for measurement.

The same failure appears in the treatment of subjective scores. The FALSE proposition that summations of subjective instrument responses are ratio measures receives 0.90 (+2.20). Even more strikingly, the FALSE proposition that summation of Likert question scores creates a ratio measure receives 0.95 (+2.50). These are among the strongest endorsements in the Norwegian profile. Assigning integers to ordered response categories and summing them produces a score; it does not demonstrate equal units, invariance, a non-arbitrary zero or ratio structure. Increasing the number of questions does not solve the problem. Nor do reliability, internal consistency, responsiveness, predictive performance or correlation with another instrument establish the measurement properties required for arithmetic. A score may be useful for description, classification or prediction without constituting a measure capable of supporting the quantitative claims attributed to it. The Norwegian profile provides little indication that this distinction functions as a binding constraint upon established HTA practice.

The weakness becomes still more pronounced when the interrogation turns to latent attributes [3]. The TRUE proposition identifying linear ratio and Rasch logit ratio measurement as the two relevant measurement pathways receives only 0.05 (−2.50). The proposition that transformation of subjective responses to interval measurement requires Rasch measurement rules receives 0.05 (−2.50). The proposition that the Rasch logit ratio scale provides the basis for assessing therapy impact for latent traits receives 0.05 (−2.50), while recognition of the correspondence between Rasch measurement requirements and the axioms of representational measurement likewise receives 0.05 (−2.50). These values occupy the floor of the normalized interrogation scale.

This absence is consequential because many attributes central to claims of therapy impact are latent. Pain, fatigue, physical function, emotional burden, treatment satisfaction and need fulfilment cannot simply be read from a physical instrument in the manner of survival time or walking distance. They require construction of a measurement system capable of locating persons and items on a defined unidimensional continuum. Without such a structure the analyst has responses, categories and scores; the analyst does not automatically have measures. The Norwegian knowledge base appears highly developed in modelling numerical outcomes while displaying near-absence of the measurement principles required to construct latent measures capable of supporting those outcomes. The TRUE proposition that the outcome of interest for latent traits is possession of the attribute receives only 0.25 (−1.10). The focus of legacy HTA therefore remains primarily upon scores, utilities and model inputs rather than the logically prior scientific problem of establishing and measuring possession of a defined attribute.

The consequences become unavoidable when the analysis moves downstream to cost-effectiveness. The TRUE proposition that claims for cost-effectiveness fail the axioms of representational measurement receives an endorsement probability of only 0.15 (−1.75). This result is entirely consistent with the rest of the profile. If the utility entering the QALY has not been established as an admissible measure for multiplication by time, the resulting QALY does not acquire measurement legitimacy merely because the multiplication has been performed. If the QALY has not been established as the required quantitative measure, subtraction does not repair the problem. An incremental QALY remains the difference between quantities whose measurement status has not been demonstrated. The subsequent division of incremental costs by incremental QALYs introduces another arithmetic operation, but division cannot repair the denominator. The expression cost per QALY does not establish that the QALY possesses the scale properties necessary to support the ratio.

This is a fundamental weakness because the ICER occupies a central position in legacy HTA. Its numerical precision can obscure the fact that its empirical meaning depends upon the measurement status of the denominator. If that denominator has not been established as a legitimate measure, greater precision in costs, more elaborate modelling, narrower uncertainty ranges or increasingly sophisticated probabilistic analysis cannot solve the problem. Arithmetic cannot rescue a failure of measurement by adding further arithmetic.

The strong endorsement of the FALSE proposition that the QALY is dimensionally homogeneous, at 0.90 (+2.20), exposes another aspect of the same failure. Dimensional homogeneity is not a cosmetic property. A purported dimensionless adjustment applied multiplicatively to time must possess the measurement structure required to support that interpretation. The designation utility does not establish dimensionlessness, and normalizing or anchoring a numerical index does not make it a legitimate proportional multiplier. If different permissible numerical representations of the underlying preference ordering generate different substantive QALY ratios, the resulting construct cannot possess the invariance required of a measure. The Norwegian knowledge base nevertheless strongly reinforces dimensional coherence at the level of the QALY while weakly possessing the principles required to demonstrate that coherence. The problem is therefore deeper than an unsuccessful proof. The requirement for the proof does not appear to function as a gatekeeper.

Norway's reputation for procedural rigor makes the results for falsifiability particularly revealing. The TRUE proposition that non-falsifiable claims should be rejected receives a moderately positive endorsement of 0.60 (+0.40), suggesting some recognition of falsifiability as a general scientific requirement. Yet the FALSE proposition that reference-case simulations generate falsifiable claims receives 0.90 (+2.20). The contradiction is important. There appears to be recognition of falsifiability in principle while the reference-case framework is simultaneously treated as though its simulated outputs satisfy that requirement. A model-generated projection conditional upon structural assumptions, extrapolated survival, utility trajectories, parameter distributions and numerous modelling choices does not become independently falsifiable merely because it is numerically explicit.

Sensitivity analysis does not resolve this distinction. It asks what happens to a model output when assumptions or parameter values are changed. Falsification asks whether an empirical claim survives confrontation with evidence. Scenario analysis, probabilistic sensitivity analysis and simulation can explore the consequences of assumptions within a model; they cannot transform a conditional imaginary projection into an independently testable therapy-impact claim. The Norwegian framework appears strongly equipped to examine uncertainty within the model while providing little evidence that empirical falsification governs the construction of the principal economic claims produced by the model.

This is where the Norwegian findings become particularly serious. Norway cannot be characterized as an HTA environment lacking organization, expertise or methodological sophistication. The opposite is true. Its national architecture is deliberately structured. Roles are defined. Evidence is assessed systematically. Models are scrutinized. Uncertainty is explored. Priority-setting criteria are explicit. Decision pathways are institutionalized. Yet none of these characteristics demonstrates that a utility is a measure, that a QALY is a ratio quantity, that QALYs are dimensionally homogeneous, that they can legitimately be aggregated or that an ICER constructed from them is scientifically meaningful.

The distinction between procedural validity and measurement validity is therefore decisive. A procedure can be followed perfectly and still operate upon inadmissible quantities. A model can be programmed correctly and still manipulate numbers that lack the scale properties required for the operations performed. A sensitivity analysis can be exhaustive and still explore uncertainty around a construct whose measurement status was never established. A national decision process can be transparent, consistent and reproducible while repeatedly reproducing the same underlying measurement failure. Procedural rigor cannot compensate for measurement failure because the two address fundamentally different questions.

Nor does institutional maturity provide the missing justification. Indeed, maturity can make the problem more difficult to identify because repeated use normalizes the constructs involved. Utilities, QALYs and ICERs become familiar. Familiarity becomes convention. Convention becomes methodological guidance. Guidance becomes training. Training produces another generation of analysts for whom the inherited arithmetic appears simply to define quantitative HTA. At no point in that sequence is measurement demonstrated. Institutional maturity demonstrates that a methodology has become established. It does not demonstrate that the quantities on which that methodology depends are measures.

This provides the distinctive interpretation of the Norwegian results. *Nye metoder* represents an attempt to manage the introduction of health technologies through a disciplined national process. Technologies are identified, assessed and appraised; evidence is assembled; economic consequences are modelled; uncertainty is considered; and decisions are made within an explicit institutional structure. Yet the interrogation indicates that the system has not imposed comparable discipline upon the measurement foundations of the quantities entering those assessments. The technology is managed, the evidence process is managed, the modelling process is managed, the uncertainty is managed and the decision pathway is managed. The measurement is not.

This is not a minor omission from an otherwise complete framework. Measurement is logically prior to the arithmetic upon which the framework depends. Failure at that point propagates through everything downstream. The QALY does not become a ratio measure because it appears in national guidance. An ICER does not become scientifically meaningful because it is produced by an approved model. A reference-case simulation does not become falsifiable because its assumptions are transparent. A preference score does not become a measure because a valuation algorithm generates it.

Nor can international acceptance provide the missing scientific justification. Norway operates within a wider legacy HTA environment in which utilities, QALYs, cost-effectiveness modelling and reference-case conventions have been normalized over decades. International inheritance can explain why these methods became embedded in Norwegian HTA. It cannot demonstrate their measurement legitimacy. Replication of a methodological convention across countries is not replication of a scientific result.

The interrogation therefore identifies something more serious than an isolated technical weakness. It reveals an epistemic architecture in which the normal order of quantitative science has been reversed. The required order is attribute, measurement, scale, admissible arithmetic, therapy-impact claim, and then empirical evaluation, replication and falsification. The Norwegian profile instead reveals a system in which arithmetic and modelling have already been institutionalized while the measurement requirements capable of determining whether that arithmetic is legitimate remain weakly represented.

The directional consistency of the findings is particularly damaging. TRUE propositions capable of constraining legacy HTA arithmetic are overwhelmingly weakly possessed. FALSE propositions necessary to sustain the established arithmetic are overwhelmingly strongly possessed. The knowledge base is therefore not merely missing isolated elements of measurement theory. Its pattern of possession and non-possession systematically favors preservation of the existing quantitative framework.

This is why the Norwegian findings should not be softened by appeals to the sophistication, transparency or maturity of the national HTA system. Those characteristics make the result more consequential, not less. Norway demonstrates how far an analytical system can develop after measurement has ceased to function as the gatekeeper of arithmetic. The result is a highly managed national HTA process resting upon numerical constructs whose required measurement properties have neither been comparably managed nor demonstrated.

The diagnosis is therefore unequivocal. Norway has successfully institutionalized the processes and arithmetic of legacy HTA while failing to institutionalize the measurement science required to justify them. It has created a disciplined system for deciding how technologies should be assessed without imposing the logically prior discipline required to determine whether the central quantitative constructs used in those assessments are measures at all.

Norway has institutionalized the arithmetic. It has not institutionalized the measurement required to justify it.

CONCLUSION

The Norwegian interrogation presents a particularly consequential example of measurement inversion because the failure occurs within one of the most deliberately organized HTA environments examined. Norway has constructed a national system intended to bring consistency, transparency and procedural discipline to the introduction of new health technologies. Assessment responsibilities are defined, evidence is reviewed systematically, economic analyses are undertaken, uncertainty is explored and decisions are embedded within an explicit institutional process. By the conventional standards of legacy HTA, this represents methodological maturity.

The results of the present interrogation demonstrate why methodological maturity cannot be equated with scientific validity [4] [5] The Norwegian knowledge base strongly possesses the arithmetic and constructs of contemporary HTA while weakly possessing the measurement principles required to determine whether that arithmetic is admissible. Measurement preceding arithmetic, the ratio requirement for multiplication, unidimensionality and the axioms of representational measurement are weakly represented. In contrast, false propositions required to sustain the established framework including treatment of the QALY as a ratio measure, acceptance of its dimensional homogeneity and aggregation, and the quantitative legitimacy assigned to preference-based scores are strongly reinforced.

This asymmetry is not a peripheral methodological deficiency. It goes directly to the scientific status of the quantitative claims generated by the system. If utility has not first been established as possessing the measurement properties required for multiplication by time, the QALY cannot acquire those properties merely because the multiplication is conventional. If the QALY has not been established as an admissible measure, subtraction cannot manufacture a legitimate incremental QALY. Division of incremental costs by that quantity cannot then manufacture a scientifically meaningful cost-per-QALY ratio. The failure occurs before the economic model begins, and nothing performed downstream can repair it.

The Norwegian results are especially revealing because the knowledge base strongly recognizes time as a ratio measure while scarcely recognizing that multiplication itself requires ratio measurement. At the same time, it strongly reinforces the false proposition that the QALY is a ratio measure. This exposes the inversion at the center of the legacy framework. One component of the calculation is correctly recognized as possessing ratio properties, but the requirement that the other component independently possess the necessary measurement properties before multiplication is undertaken has effectively disappeared. The arithmetic has become the starting point rather than the consequence of measurement.

The near-absence of Rasch measurement principles extends the problem beyond the QALY. Many attributes through which therapy impact is expressed are latent. Pain, fatigue, physical function, need fulfilment and related attributes cannot acquire quantitative status merely through numerical coding, summation or statistical transformation. They require construction of a defined unidimensional measurement continuum and demonstration of the measurement properties necessary for quantitative claims. Without this, the system possesses scores and numerical representations rather than measures. The Norwegian interrogation indicates that this distinction has not been incorporated as a fundamental requirement of the national HTA knowledge environment.

The findings for falsifiability reinforce the same conclusion. Norway shows some recognition of the general proposition that non-falsifiable claims should be rejected while strongly reinforcing the false proposition that reference-case simulations generate falsifiable claims. This distinction matters. Sensitivity analysis tests the consequences of varying assumptions within a model. It does not constitute empirical falsification of the therapy-impact claim itself. A simulation can be transparent, internally consistent and computationally sophisticated without producing a claim capable of independent empirical evaluation, replication and falsification.

Norway therefore demonstrates the limitations of procedural rigor with unusual clarity. A technology-assessment process can be highly organized while the measurement foundations of its quantitative claims remain unestablished. Models can be scrutinized without scrutinizing whether their inputs are measures. Uncertainty can be explored around quantities whose scale properties have never been demonstrated. Decision rules can be applied consistently to numerical outputs whose empirical meaning remains unresolved. Reproducibility of procedure cannot substitute for validity of measurement.

This distinction should fundamentally change how the Norwegian results are interpreted. The problem is not that Norway requires additional methodological sophistication. Adding further modelling techniques, improving probabilistic analysis, refining utility estimation or increasing the complexity of economic evaluation would leave the prior failure untouched. The problem is the order in which quantitative science has been constructed. Measurement has been subordinated to an inherited arithmetic framework when it should have determined the legitimacy of that framework from the outset.

The international inheritance of legacy HTA explains how this occurred. Norway did not independently invent the QALY, cost-per-QALY analysis or the reference-case modelling paradigm. It adopted and developed methods already normalized within international HTA. That history explains institutionalization; it does not establish scientific legitimacy. Agreement among HTA agencies, professional organizations, academic centers and methodological authorities cannot confer measurement properties upon a numerical construct. Consensus can establish a convention. It cannot establish a scale of measurement.

The responsibility facing Norwegian HTA is therefore prospective. The measurement failure has now been made explicit. Continuing to apply the inherited framework as though the measurement question had never been raised would be fundamentally different from having inherited that framework before the problem was exposed. Historical transmission can explain the legacy. It cannot provide a continuing scientific justification for it.

The required transition does not demand invention of a new standard of quantitative science. The relevant standards already exist. The sequence must simply be restored to its proper order: identify the attribute, establish measurement, determine the scale and admissible transformations, undertake only the arithmetic permitted by that scale, construct the therapy-impact claim, and subject that claim to empirical evaluation, replication and falsification. Manifest attributes require appropriate linear ratio measurement. Latent attributes require construction of an invariant measurement continuum capable of establishing possession and change in the defined attribute. Arithmetic follows measurement; it does not create it.

Norway may be unusually well positioned to undertake such a transition. The institutional capacity that produced a coordinated national HTA process could also be used to impose measurement discipline upon that process. A system capable of defining assessment pathways, evidentiary requirements, methodological responsibilities and decision criteria is equally capable of requiring that quantitative claims first demonstrate their measurement foundations. The organizational achievement represented by *Nye metoder* need not therefore be abandoned. The scientific foundations upon which its quantitative component rests must be reconstructed.

That is the choice exposed by the Norwegian interrogation. Norway can continue refining a highly managed legacy HTA system in which the arithmetic is institutionally secure while its measurement foundations remain unestablished, or it can require that measurement become the gatekeeper of quantitative therapy evaluation.

The distinction is fundamental. A managed process is not necessarily a scientifically measured process. Norway has successfully institutionalized the arithmetic of legacy HTA. The next task is more important: to institutionalize the measurement required to justify it**.** Measurement first. Arithmetic second.

ACKNOWLEDGEMENT

I acknowledge that I have used OpenAI technologies, including the large language model, to assist in the development of this work. All final decisions, interpretations, and responsibilities for the content rest solely with me.

REFERENCES

---

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

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

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

[4] Langley P. California: Legacy HTA – 40 years of numerical storytelling. Logit Working Paper No 6765. https://maimonresearch.com/logit-working-paper-no-6765-august-2026/

[5] Langley P. Legacy HTA: Is this the worst we could have done? Logit Working Paper No 7987 https://maimonresearch.com/logit-working-paper-no-7987-august-2026/