MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

THE PATH TO EQ-5D HTA CLOSURE: FROM TTO TO THE IMPOSSIBLE QALY

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

LOGIT WORKING PAPER No 1818 JUNE 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

This paper examines the measurement foundations of the EQ-5D-3L and EQ-5D-5L utility frameworks and their central role in contemporary health technology assessment (HTA). The argument is straightforward: arithmetic operations are admissible only when applied to lawful measures. The history of science demonstrates that measurement precedes arithmetic. Quantitative claims require attributes that satisfy the standards of representational measurement before mathematical operations can be undertaken. In HTA, however, arithmetic has come to precede measurement.

The analysis begins with the unique position of the ratio measure. Drawing on Stevens' theory of measurement scales and the axioms of representational measurement, it is argued that quantitative claims require either linear ratio measures for manifest attributes or Rasch logit ratio measures for latent attributes. These represent the only lawful measurement frameworks capable of supporting evaluable, replicable and falsifiable claims regarding therapy impact.

Against this background, the paper traces the sequential construction of EQ-5D utilities through eight stages. The process begins with time trade-off (TTO) valuations of multidimensional health-state descriptions. These preference scores become the dependent variable in econometric models employing dummy variables to estimate average conditional effects. The resulting coefficients are subsequently transformed into decrement weights, combined through additive algorithms to create utility scores, and finally multiplied by time to generate quality-adjusted life years (QALYs).

At each stage the same question is asked: do the quantities involved satisfy the requirements of ratio measurement? The conclusion is that they do not. TTO valuations represent preferences rather than demonstrated measures. Econometric models explain variation in these valuations but cannot create measurement properties absent from the dependent variable. Regression coefficients therefore inherit the limitations of the TTO values from which they are derived. Their subsequent reinterpretation as decrement weights and their arithmetic aggregation into utility scores are unsupported by evidence that a common measurable attribute exists. The resulting utility scores are scoring-system outputs rather than demonstrated measures. Consequently, their use as discount factors for time in the construction of QALYs is inadmissible.

The broader implication is that the reference-case framework underpinning modern HTA lacks a lawful measurement foundation. Utilities, QALYs and simulation models represent successive transformations of quantities whose measurement properties remain unproven. The paper concludes that HTA can recover scientific credibility only by abandoning arithmetic-based utility constructions and returning to measurement. Therapy impact claims should instead be based on linear ratio measures for manifest attributes and Rasch logit ratio measures for latent attributes, with all claims subject to empirical evaluation, replication and falsification.

INTRODUCTION

The history of science is inseparable from the history of measurement. Since the Scientific Revolution, quantitative inquiry has depended upon the ability to identify attributes, construct measures for those attributes and evaluate claims through observation and experiment [1]. Measurement is therefore not an optional component of science. It is its foundation. Arithmetic follows measurement. It does not create it.

This principle is reflected in the scales of measurement first formalized by Stevens and later refined through the axioms of representational measurement [2 3] . Of the recognized scales, only ratio measures support the full range of arithmetic operations. A ratio measure requires a clearly defined attribute, unidimensionality, a meaningful zero, meaningful ratio comparisons and invariance. Most importantly, the numerical representation must preserve the empirical structure of the attribute being measured. Only when these conditions are satisfied can arithmetic operations such as multiplication and division be regarded as admissible.

Health technology assessment (HTA) has largely ignored these requirements. Instead, the discipline has become committed to a sequence of numerical constructions culminating in utilities, quality-adjusted life years (QALYs) and cost-effectiveness claims. The underlying assumption is straightforward. If a number can be attached to a health state, then that number can be manipulated through arithmetic operations to create increasingly sophisticated quantitative outputs. Yet the existence of a number does not establish the existence of a measure. Before arithmetic can begin, the measurement properties of the quantities involved must be demonstrated.

The EQ-5D-3L and EQ-5D-5L provide perhaps the most influential examples of this process. Health-state descriptions are first valued through time trade-off (TTO) exercises. These values then become the dependent variable in econometric models from which decrement weights are estimated. The resulting coefficients are combined through additive algorithms to create utility scores which are subsequently multiplied by time to generate QALYs. Each step involves an arithmetic operation. At no point, however, is it demonstrated that the quantities entering these calculations satisfy the requirements of ratio measurement.

The purpose of this paper is to examine this sequence in detail. The argument is straightforward. The issue is not whether the econometric models are statistically sophisticated, whether the algorithms are widely accepted or whether the resulting utility scores appear plausible. The issue is whether the arithmetic operations that define the EQ-5D-3L and EQ-5D-5L are supported by lawful measurement. If ratio measurement cannot be demonstrated, then the subsequent arithmetic becomes inadmissible. The result is measurement inversion: arithmetic preceding measurement. The EQ-5D utility framework therefore provides a revealing example of how HTA has substituted numerical manipulation for measurement and, in doing so, abandoned the standards that govern every other quantitative science.

THE UNIQUE POSITION OF THE RATIO MEASURE

At the center of every quantitative science lies the concept of measurement. Without measurement there can be no meaningful arithmetic, no quantitative comparison and no evaluable scientific

claims. Since the Scientific Revolution, scientific progress has depended upon the ability to identify attributes, construct lawful measures for those attributes and evaluate claims through observation and experiment. Measurement therefore precedes arithmetic. It does not emerge from arithmetic.

The importance of this principle was formalized by Stevens in his classic 1946 paper on scales of measurement. Stevens distinguished four scale types: nominal, ordinal, interval and ratio. These scales describe the ways in which observations of manifest attributes can be represented numerically. Nominal scales classify observations. Ordinal scales rank observations. Interval scales permit meaningful comparisons of differences but lack a meaningful zero. Ratio scales alone possess all required properties, including a meaningful non-arbitrary zero that permits proportional comparisons. Only ratio scales support the full range of arithmetic operations, including multiplication and division.

The ratio scale therefore occupies a unique position. A ratio measure requires a clearly defined attribute, unidimensionality, a meaningful zero, meaningful ratio comparisons and invariance. Most importantly, the numerical representation must preserve the empirical structure of the attribute being measured. These requirements were subsequently formalized through the axioms of representational measurement developed by Krantz, Luce, Suppes and Tversky in 1971. Measurement is not merely the assignment of numbers. It is the construction of a lawful numerical representation of an attribute.

Stevens' framework, however, addresses only manifest attributes: attributes that can be observed directly. Length, weight, survival time, hospital days, physician visits and laboratory values are examples. For such attributes the objective is to construct a linear ratio measure capable of supporting quantitative claims and empirical evaluation.

The situation is different for latent attributes. Pain, fatigue, depression, anxiety, physical functioning and need fulfilment cannot be observed directly. They must be inferred from patterns of observable responses. For many decades this distinction was poorly understood, leading to the widespread use of ordinal scores, summed scales and composite indices as though they were measures. The solution to this problem was provided by Rasch measurement theory in 1960 [4] .

Rasch demonstrated that latent attributes can be measured provided observations satisfy a set of measurement requirements equivalent in purpose to the axioms of representational measurement [5] . The result is not a linear ratio scale but a Rasch logit ratio scale, where the possession of a latent attribute is represented through the logarithm of an odds ratio. This provides the only framework capable of transforming observations into lawful measures of latent attributes while preserving invariance and specific objectivity.

The implications for health technology assessment are straightforward. All claims for therapy impact refer either to manifest attributes or latent attributes. Manifest attributes require linear ratio measures. Latent attributes require Rasch logit ratio measures. There are no other measurement possibilities. Once this distinction is recognized, the measurement problem in HTA becomes remarkably simple. The challenge is not to create utilities, QALYs or composite indices. The

challenge is to identify the attribute of interest and construct a lawful measure appropriate to its type.

The ratio measure therefore occupies a unique position in HTA. Whether expressed as a linear ratio measure for manifest attributes or as a Rasch logit ratio measure for latent attributes, it is the indispensable foundation for quantitative claims. Without ratio measurement, arithmetic becomes inadmissible and numerical constructions become indistinguishable from numerical storytelling. With ratio measurement, therapy impact claims become evaluable, replicable and falsifiable. Measurement is not one component of HTA. It is its foundation.

THE FIRST STEP TO CLOSURE: THE TIME TRADE-OFF

The foundation of the EQ-5D utility framework is the time trade-off (TTO) technique. Every subsequent stage in the construction of utilities and QALYs depends upon the values generated by this procedure. The econometric model, the decrement weights, the utility algorithm and the QALY itself are all derived from TTO valuations. Consequently, if the TTO fails to satisfy the requirements of measurement, every subsequent arithmetic operation inherits the same deficiency. The critical question is therefore not whether the TTO generates numbers. It clearly does. The question is whether those numbers constitute lawful measures.

The TTO is designed to elicit preferences for hypothetical health states. Respondents are presented with a health-state description, typically defined by the five EQ-5D dimensions of mobility, self-care, usual activities, pain/discomfort and anxiety/depression. They are then asked to choose between a specified period of time in that health state and a shorter period of time in full health. The point of indifference between the two alternatives generates a numerical value. If ten years in the health state is judged equivalent to five years in full health, the resulting TTO value is 0.5. These values are commonly interpreted as utilities and subsequently become the dependent variable in econometric estimation.

The crucial issue is that the TTO does not demonstrate measurement. It demonstrates preference. The respondent is not measuring a health attribute. Rather, the respondent is expressing a judgment regarding the desirability of a hypothetical health-state description. The resulting numerical value is therefore a preference score. Whether that score possesses the characteristics of a ratio measure is simply assumed.

This assumption is particularly problematic because the health-state descriptions themselves are multidimensional. A respondent valuing health state 21123 is simultaneously considering mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The resulting valuation therefore reflects a composite judgment regarding multiple dimensions of health. Yet lawful measurement begins with the identification of a single attribute. Without unidimensionality there can be no ratio measurement. The TTO provides no evidence that respondents are valuing a single attribute rather than forming an overall preference regarding a multidimensional health-state description.

Nor does the TTO establish the other requirements of ratio measurement. There is no demonstration of a meaningful non-arbitrary zero. There is no demonstration that proportional

comparisons are meaningful. There is no evidence that a valuation of 0.8 represents twice the quantity represented by a valuation of 0.4. There is no demonstration that the numerical assignments preserve the empirical structure of a measurable attribute. Most importantly, there is no attempt to satisfy the axioms of representational measurement that define the conditions under which arithmetic operations become admissible.

The distinction between preference and measurement is fundamental. Preferences can be expressed numerically without creating measures. Assigning numbers to judgments does not establish the existence of a quantitative attribute. Yet the TTO framework proceeds as though numerical preference scores are synonymous with measurement.

The consequence is profound. The TTO generates a collection of respondent valuations attached to hypothetical health-state descriptions. These valuations may be useful as indicators of preference, but their measurement properties remain unknown. Nevertheless, they become the foundation for every subsequent stage of utility construction. The econometric model does not create new measurement properties. The decrement weights do not create new measurement properties. The utility algorithm does not create new measurement properties. All are mathematical transformations of the original TTO values.

The first step to HTA closure therefore occurs at the very beginning. Preference is mistaken for measurement. Numerical valuations are accepted as though they were lawful measures. Once this assumption is accepted, the entire chain of utility construction follows. If ratio measurement is absent at the start, it remains absent throughout.

THE SECOND STEP TO CLOSURE THE CHOICE OF DEPENDENT VARIABLE

Having accepted the TTO valuation as the foundation of the utility framework, the next step is to employ that valuation as the dependent variable in an econometric model. At first glance this appears to be an unremarkable statistical exercise. Regression models require a dependent variable, and the TTO provides one. Yet from the perspective of measurement this step is critical because it reveals a fundamental misunderstanding. Regression analysis assumes measurement. It does not create it.

The nature of the dependent variable deserves careful attention. The dependent variable is not health status, quality of life, mobility, pain, anxiety or any other identifiable health attribute. Rather, each observation in the dataset represents a respondent's valuation of a hypothetical health-state description. If respondent (i) values health state (j), then the dependent variable Y_{ij} is the TTO valuation assigned to that health state.

The econometric dataset therefore consists of a large number of respondent-health-state combinations. A single health state may be valued by many respondents. Each respondent may value several health states. The result is a dataset containing hundreds or thousands of observations of TTO valuations. The crucial point, however, is that increasing the number of observations does not alter the nature of the dependent variable. The dependent variable remains a preference-based valuation attached to a hypothetical health-state description.

This distinction is fundamental because the econometric model is often treated as though it somehow transforms these valuations into measures. It does not. The model simply seeks to explain variation in the TTO scores. It asks why one health-state description receives a higher average valuation than another. It does not ask whether the valuations themselves possess the properties required for measurement.

The issue becomes even clearer when the requirements of representational measurement are considered. Before a variable can support quantitative inference, it must refer to a clearly defined attribute. That attribute must satisfy the requirements of measurement. Yet the dependent variable in the TTO model is not a measured attribute. It is a respondent's valuation of a multidimensional health-state description. The model therefore explains variation in preferences rather than variation in a demonstrated quantitative attribute.

This distinction is often obscured because econometric estimation produces numerical coefficients and statistical tests. The appearance of mathematical rigor can create the impression that measurement has been established. In reality, the model merely decomposes variation in the dependent variable. Whatever measurement properties the dependent variable possesses are inherited by the resulting coefficients. The regression cannot create stronger measurement properties than those already present in the dependent variable itself.

This observation has profound implications. If the TTO valuation has not been demonstrated to be a lawful ratio measure, then the regression coefficients cannot be ratio measures merely because they emerge from a statistical model. They remain estimates expressed in the units of the dependent variable. The econometric procedure may improve prediction. It may improve fit. It may explain variation. What it cannot do is transform preference scores into lawful measures.

The consequence is that the utility framework commits a second act of measurement inversion. Rather than first establishing the measurement status of the dependent variable and then proceeding to statistical analysis, it proceeds directly to econometric estimation while leaving the measurement question unresolved. The model therefore rests upon an assumption that remains unproven: that a respondent's valuation of a hypothetical health-state description constitutes a lawful measure.

This is the second step to HTA closure. The TTO valuation is accepted as the dependent variable without demonstrating its measurement properties. The econometric model then treats that valuation as though it were a quantitative measure capable of supporting statistical decomposition. Yet the model inherits the limitations of the dependent variable. It cannot transcend them. Whatever deficiencies exist in the TTO valuation are carried forward in

THE THIRD STEP TO CLOSURE: THE DUMMY VARIABLE MODEL

Once the TTO valuation has been accepted as the dependent variable, the next stage in the construction of the EQ-5D tariff is the specification of an econometric model based upon dummy variables. This stage is often presented as a technical exercise in statistical estimation, but its implications for measurement are rarely examined. The issue is not whether the use of dummy

variables is statistically legitimate. It unquestionably is. The issue is whether the resulting model contributes anything toward establishing measurement.

The answer is that it does not.

The EQ-5D descriptive system defines health states according to five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. In the three-level version of the instrument, each dimension has three levels. Level 1 represents no problems and serves as the reference category. Levels 2 and 3 are represented through dummy variables. The econometric model therefore contains ten dummy variables corresponding to the non-reference levels of the five dimensions.

The important point is that these dummy variables are not measures. They are classifications. A dummy variable merely records whether a particular category is present or absent within a health-state description. For example, the presence of Mobility Level 2 is recorded as 1 and its absence as 0. The same principle applies to every other dimension and level.

This distinction is critical because it is often assumed that the dummy variables somehow represent quantities of health. They do not. The dummy variables simply describe the structure of the health-state description being valued. They identify categories. They do not measure attributes.

The health state 21123, for example, automatically determines which dummy variables are present. The econometric model does not attempt to discover these categories. They are known with certainty before estimation begins. The purpose of the regression is therefore not to measure health dimensions but to estimate the average effect of each category on the TTO valuation.

This observation has important consequences for the interpretation of the coefficients. The coefficient associated with Pain Level 2, for example, does not measure pain. Nor does it measure utility. It represents the estimated average change in the TTO valuation associated with the presence of Pain Level 2 relative to Pain Level 1. Similar interpretations apply to every other coefficient in the model.

The distinction between measurement and association is fundamental. The model estimates relationships between health-state categories and preference scores. It does not establish that any of the categories correspond to quantitative attributes. Nor does it establish that the resulting coefficients are measures. The regression merely describes how respondents' valuations change when particular categories are present within a health-state description.

A further complication arises because the coefficients are estimated in the presence of multiple dimensions simultaneously. Respondents do not value mobility, pain or anxiety in isolation. They value complete health-state descriptions. Consequently, each coefficient represents an average conditional effect estimated across many combinations of the remaining dimensions. The coefficients therefore do not represent independent quantities. They represent statistical summaries derived from a multidimensional specification.

This point is often overlooked when the coefficients are subsequently interpreted as decrement weights. The econometric model provides estimates of average conditional effects on TTO valuations. It does not provide measurements of health dimensions. Yet the distinction is rarely acknowledged.

This is the third step to HTA closure. The utility framework moves from preference scores to a dummy-variable model and creates the appearance of quantitative analysis. The model is statistically conventional, but it contributes nothing toward establishing measurement. The dummy variables are classifications rather than measures, and the coefficients are average conditional effects rather than quantities of health. Arithmetic advances once again, while the question of measurement remains unanswered.

THE FOURTH STEP TO CLOSURE: FROM CONDITIONAL EFFECTS TO QUANTITATIVE CLAIMS

The econometric model produces a set of coefficients that are commonly interpreted as the contribution of each EQ-5D dimension and level to the utility score. This interpretation appears natural. The model has been estimated, the coefficients are statistically significant, and numerical values have been obtained. Yet this stage represents one of the most important conceptual errors in the entire utility construction process. The coefficients are not measures. They are average conditional effects on the dependent variable.

This distinction is crucial.

Consider a coefficient associated with Pain/Discomfort Level 2. Suppose the estimated coefficient is -0.067. The meaning of this coefficient is straightforward. Across the valuation dataset, health states containing Pain Level 2 are associated with average TTO valuations that are 0.067 lower than comparable health states with Pain Level 1, all else being equal within the model specification. This is a statistical statement. It describes an estimated relationship between a category and a preference score.

What it does not do is establish measurement.

The coefficient is not a measure of pain. It is not a measure of health loss. It is not a measure of utility. It is not even a measure of the contribution of pain to an underlying health attribute. It is merely an estimate of the average change in the dependent variable associated with the presence of a particular category.

The importance of this observation cannot be overstated because the coefficient inherits the properties of the dependent variable from which it is estimated. If the dependent variable is a TTO valuation whose measurement properties have not been established, then the coefficient cannot somehow acquire stronger measurement properties through statistical estimation. Regression analysis decomposes variation in the dependent variable. It does not transform preference scores into measures.

This principle is well understood in econometrics. A regression coefficient is expressed in the units of the dependent variable. If the dependent variable is dollars, the coefficient is expressed in dollars. If the dependent variable is years, the coefficient is expressed in years. Similarly, if the dependent variable is a TTO valuation, the coefficient is expressed in the units of that valuation. The coefficient cannot possess measurement properties that are absent from the dependent variable itself.

This creates a profound problem for the utility framework. The coefficients are frequently discussed as though they were quantitative decrements representing amounts of health lost. Yet the regression model establishes no such interpretation. It merely estimates average conditional effects on preference scores. The transition from statistical parameter to quantitative decrement is assumed rather than demonstrated.

The problem becomes even more acute when interactions are considered. The coefficients are not estimated from respondents valuing individual dimensions in isolation. Respondents value complete health-state descriptions containing multiple dimensions simultaneously. Consequently, each coefficient reflects an average effect across numerous combinations of mobility, self-care, usual activities, pain and anxiety. The coefficient is therefore model-dependent. It is a statistical summary of complex interactions within the valuation dataset rather than a measurement of an independent quantity.

This is the fourth step to HTA closure. Average conditional effects become transformed into apparent quantities. Statistical parameters begin to acquire the language of measurement despite the absence of any demonstration that measurement has occurred. The econometric model has done precisely what it was designed to do: explain variation in preference scores. What it has not done is establish that the resulting coefficients are measures. Yet it is precisely this unwarranted transformation that makes the subsequent construction of utility weights possible. The journey from preference to arithmetic continues, while measurement remains conspicuously absent.

THE FIFTH STEP TO CLOSURE: FROM CONDITIONAL EFFECTS TO DECREMENT WEIGHTS

The next stage in the construction of the EQ-5D utility framework is perhaps the most remarkable. The econometric model has produced a series of coefficients representing average conditional effects on TTO valuations. At this point the coefficients are simply statistical parameters. They describe average changes in the dependent variable associated with particular health-state categories. Yet in the utility construction process these coefficients undergo a profound transformation. They cease to be statistical effects and become decrement weights.

This transformation lies at the heart of the utility algorithm.

Consider again the coefficient associated with Pain/Discomfort Level 2. Suppose the econometric model estimates a value of -0.067. The statistical interpretation is straightforward. Health states containing Pain Level 2 are associated with average TTO valuations that are 0.067 lower than otherwise comparable health states with Pain Level 1. The coefficient therefore summarizes an average conditional relationship within the valuation dataset.

The utility framework assigns a very different interpretation. The coefficient is detached from the regression model and becomes a decrement weight attached to the questionnaire response category itself. Whenever a respondent reports Pain Level 2, the utility algorithm applies the decrement of 0.067. The coefficient is no longer treated as an average conditional effect estimated from a population valuation exercise. It is treated as though it were a quantity of health loss associated with that response category.

This is a conceptual leap of extraordinary importance.

The regression model did not estimate quantities of health loss. It estimated changes in a dependent variable. The coefficient is therefore a statistical summary of observed relationships in the valuation dataset. Nothing in the estimation process demonstrates that the coefficient represents a measurable quantity. Nothing demonstrates that it is a ratio measure. Nothing demonstrates that it can legitimately be interpreted as a decrement in a common health attribute.

Yet this reinterpretation is essential to the utility framework. Without it there can be no utility algorithm. The coefficients must be transformed into decrement weights because the next stage of the process requires numbers that can be combined arithmetically. The statistical parameters therefore acquire a new identity. They become building blocks in the construction of a utility score.

The difficulty is that the measurement status of these weights remains entirely unknown. If the TTO valuation is not a demonstrated ratio measure, then the coefficients estimated from that valuation cannot be ratio measures. The coefficients inherit the properties of the dependent variable and nothing more. Consequently, the decrement weights are not established measures. They are numerical transformations of numerical transformations. Preference scores become regression coefficients, and regression coefficients become decrement weights.

The problem is compounded by the fact that each coefficient is estimated as an average conditional effect across many combinations of health dimensions. A coefficient attached to Pain Level 2 reflects not only pain but the context in which that category appeared during estimation. It is therefore model-dependent. Nevertheless, the utility framework subsequently treats the coefficient as a fixed quantity that can be applied uniformly to every respondent who selects that response category.

This is the fifth step to HTA closure. Statistical parameters become transformed into apparent quantities without demonstrating that measurement has occurred. The average conditional effects estimated by the regression model are reinterpreted as decrement weights attached to questionnaire responses. The transition appears natural because the coefficients are numerical. Yet numbers are not measures simply because they are numerical. The utility framework quietly assumes that statistical effects can become quantities. It is this assumption that permits the next and even more problematic stage: the arithmetic combination of decrement weights to create a utility score.

THE SIXTH STEP TO CLOSURE: THE ADDITION OF DECREMENT WEIGHTS

Having transformed average conditional effects into decrement weights, the utility framework now undertakes its most ambitious step. The individual decrement weights are combined arithmetically to create a single utility score. This stage is so familiar within health economics that it is rarely questioned. Yet from the perspective of measurement theory it represents one of the most problematic assumptions in the entire construction process.

The issue is not the arithmetic itself. Addition is straightforward. The issue is whether the quantities being added possess the properties required for meaningful addition.

Consider a respondent who reports health state 21123. The utility algorithm identifies the corresponding decrement weights attached to the relevant response categories. These weights are then combined according to the scoring formula. The result is a single utility score that is subsequently interpreted as representing the respondent's overall health status relative to full health.

At first sight this appears entirely reasonable. If each dimension contributes a decrement, why should the decrements not be added together? The answer lies in the requirements of measurement.

Addition is not a purely mathematical operation. It is a measurement operation. For quantities to be added meaningfully they must refer to the same underlying attribute. This requirement is often described as dimensional homogeneity. Lengths can be added to lengths. Weights can be added to weights. Time intervals can be added to time intervals. The reason is simple: each quantity represents a different amount of the same attribute.

The utility algorithm assumes precisely this condition. It assumes that the decrement associated with mobility, the decrement associated with pain, the decrement associated with anxiety, the decrement associated with self-care and the decrement associated with usual activities are all quantities of a common attribute that can be combined into a single total.

Yet nowhere is this assumption demonstrated.

The EQ-5D dimensions were never constructed as measurements of a single attribute. They are distinct aspects of health. Mobility is not pain. Anxiety is not self-care. Usual activities are not mobility. The econometric model estimates average conditional effects associated with these categories, but it does not demonstrate that they represent quantities of a common measurable attribute.

The problem is deeper still. The decrement weights themselves are not measurements. They are transformed regression coefficients derived from TTO valuations. As argued previously, these coefficients inherit whatever measurement properties the TTO valuations possess. If the TTO valuations are not demonstrated ratio measures, then the decrement weights cannot be ratio measures. The addition of such weights therefore cannot create measurement where measurement was absent at the outset.

Indeed, the utility framework performs two acts of faith simultaneously. First, it assumes that the decrement weights represent quantitative entities. Second, it assumes that these entities belong to a common dimension that permits addition. Neither assumption is supported by evidence. Both are simply embedded within the scoring algorithm.

The consequences are profound. The resulting utility score appears to be a measure because it is expressed as a single number. Yet numerical aggregation is not measurement. A single number can always be produced by combining other numbers. The existence of a numerical result does not establish that the result measures anything.

This is the sixth step to HTA closure. Decrement weights derived from average conditional effects are combined as though they represented quantities of a common attribute. Dimensional homogeneity is assumed rather than demonstrated. Arithmetic substitutes for measurement. The resulting utility score is therefore not the outcome of a measurement process but the outcome of a scoring process. The distinction is crucial because the entire utility framework depends upon treating this score as though it were a lawful measure. The next step is therefore inevitable: the interpretation of the resulting score as a utility.

THE SEVENTH STEP TO CLOSURE: THE UTILITY SCORE AS AN APPARENT MEASURE

The culmination of the utility construction process is the creation of the utility score itself. After preference valuations have been obtained through the time trade-off, after econometric models have estimated average conditional effects, after those effects have been transformed into decrement weights, and after the weights have been combined through arithmetic operations, the result is a single numerical value. This value is then interpreted as a utility. It is at this point that the distinction between a scoring system and a measurement system becomes critically important.

The utility score appears to possess all the characteristics of a quantitative measure. It is expressed as a single number. It is bounded by convention, typically between zero and one, although negative values may also be permitted. It can be compared across individuals and populations. It can be manipulated arithmetically and ultimately multiplied by time to generate QALYs. The appearance of measurement is therefore compelling.

The appearance, however, should not be mistaken for reality.

The utility score is not the result of a measurement process. It is the result of a scoring process. The distinction is fundamental. Measurement begins with an attribute and demonstrates that the numerical assignments preserve the empirical structure of that attribute. Scoring begins with a set of numerical rules and produces a numerical result. The existence of a score does not establish the existence of a measure.

This distinction is easily overlooked because the utility score is presented as though it represents a quantity of health. Yet no such quantity has ever been demonstrated. The score emerges from the arithmetic aggregation of decrement weights, which themselves are transformed regression coefficients derived from TTO valuations. At no stage has a single unidimensional attribute been

identified and measured. At no stage have the requirements of representational measurement been satisfied. At no stage has the score been shown to possess the properties required of a ratio measure.

The problem is therefore cumulative. Every uncertainty associated with the TTO valuation remains embedded in the final utility score. Every uncertainty associated with the regression coefficients remains embedded in the final utility score. Every uncertainty associated with the decrement weights remains embedded in the final utility score. The utility algorithm merely combines these uncertainties into a single numerical output.

This observation leads to an important principle. A mathematical transformation cannot create measurement properties that were absent in the original data. The utility score is a function of the TTO valuations from which it ultimately derives. Consequently, the utility score cannot possess stronger measurement properties than those valuations. If the TTO valuations are not demonstrated ratio measures, then neither the utility score nor any subsequent construct derived from it can legitimately claim ratio-scale status.

The bounded nature of the utility score does not solve this problem. A number expressed between zero and one is not automatically a proportion. A value of 0.8 can only be interpreted as eighty percent of a quantity if the scale possesses lawful ratio properties. Similarly, a value of 0.4 can only be interpreted as half of 0.8 if proportional comparisons are meaningful. The utility framework assumes these properties. It does not demonstrate them.

This is the seventh step to HTA closure. A scoring algorithm produces a numerical result and that result is reclassified as a utility. The transformation appears innocuous because the output is numerical. Yet numerical outputs are not necessarily measures. The utility score is best understood as the final product of a sequence of arithmetic operations applied to preference-based valuations. It is a score generated by an algorithm. Whether it is a measure remains entirely unproven. Nevertheless, the utility framework proceeds as though that question has already been answered.

THE EIGHTH STEP TO PERDITION: THE IMPOSSIBLE QALY

The final stage in the utility framework is the construction of the quality-adjusted life year (QALY). At this point the utility score, itself the product of a lengthy sequence of preference elicitation, econometric estimation and arithmetic aggregation, is multiplied by time to generate a measure of health gain. The QALY is widely regarded as the principal outcome of cost-effectiveness analysis and the cornerstone of modern health technology assessment. Yet from the perspective of measurement theory, the QALY represents the culmination of every unresolved problem encountered in the preceding stages.

Time presents no difficulty. Time is a lawful ratio measure. It possesses a meaningful non-arbitrary zero. Ratio comparisons are meaningful. Twelve months is twice six months. Twenty-four months is twice twelve months. The measurement properties of time are well established and universally accepted.

The entire legitimacy of the QALY therefore rests upon the utility score.

The utility is treated as a discount factor applied to time. If a respondent has a utility of 0.5 and remains in that health state for twelve months, the resulting QALY is 0.5. This is interpreted as six months in perfect health. Similarly, a utility of 0.8 over twelve months is interpreted as 9.6 months in perfect health. These are not merely arithmetic calculations. They are substantive quantitative claims. They assume that the utility score possesses the properties required of a ratio measure.

This assumption is unavoidable. A discount factor capable of converting twelve months into six months of perfect health must itself be a lawful ratio-scale quantity. There must be a clearly defined attribute. The attribute must be unidimensional. The scale must possess a meaningful zero. Proportional comparisons must be meaningful. The numerical representation must satisfy the requirements of representational measurement. Without these properties, multiplication becomes an exercise in arithmetic rather than measurement.

The difficulty is that none of these requirements have been demonstrated for the utility score. The utility is not derived from the measurement of a single attribute. It emerges from the aggregation of coefficients attached to mobility, self-care, usual activities, pain/discomfort and anxiety/depression. These are distinct dimensions of health. The utility score therefore originates in a multidimensional classification system rather than a unidimensional attribute. No evidence is provided that the resulting score represents a single measurable quantity. No evidence is provided that proportional comparisons are meaningful. No evidence is provided that the score satisfies the requirements of ratio measurement.

Indeed, the utility score inherits every unresolved problem associated with the original TTO valuation. The econometric model did not create measurement. The regression coefficients did not create measurement. The decrement weights did not create measurement. The utility algorithm did not create measurement. Each stage merely transformed the output of the preceding stage. Consequently, the utility score cannot possess stronger measurement properties than the TTO valuations from which it ultimately derives.

The implication is profound. If the utility score is not a lawful ratio measure, then it cannot function as a discount factor for time. If it cannot function as a discount factor for time, then the multiplication at the heart of the QALY is inadmissible. The resulting QALY is not a measure of health gain. It is the product of multiplying a ratio measure of time by a numerical score whose measurement properties remain unknown.

This is the final step to HTA closure. Preference scores become regression coefficients. Regression coefficients become decrement weights. Decrement weights become utility scores. Utility scores become discount factors for time. At every stage arithmetic advances while measurement retreats. The QALY therefore stands not as the triumph of quantitative assessment but as its inversion. The impossible QALY is the inevitable consequence of accepting arithmetic in place of measurement.

IMPLICATIONS FOR HTA: DOES THE REFERENCE CASE HAVE A FUTURE?

The implications of this analysis extend far beyond the construction of EQ-5D utilities and QALYs. The central question is whether health technology assessment, as currently practiced, can

continue to rely upon the reference-case framework that has dominated the field for more than four decades. The answer depends upon whether the reference case can satisfy the standards of measurement required for quantitative science.

The reference case rests upon a simple foundation. Utility scores are assumed to be valid measures of health-related quality of life. These utility scores are multiplied by time to create QALYs. QALYs are then incorporated into simulation models that generate estimates of lifetime costs, lifetime benefits and incremental cost-effectiveness ratios. The resulting projections become the basis for reimbursement recommendations, pricing decisions and resource allocation.

The difficulty is that every stage of this process depends upon the validity of the utility score. If utilities fail the requirements of representational measurement, then the QALY fails. If the QALY fails, then the simulation model fails. If the simulation model fails, then the reference case loses its quantitative foundation.

This is not a criticism of simulation modelling as such. Simulation models can be useful tools when they operate on lawful measures. The problem is that the reference case begins with quantities whose measurement status has never been demonstrated. The model therefore inherits the deficiencies of its inputs. Sophisticated mathematics cannot rescue invalid measurement. A simulation model can only transform its inputs. It cannot create measurement properties that were absent at the outset.

The consequence is that the reference case resembles a system of numerical storytelling rather than empirical science. Claims regarding lifetime health gains, future costs and incremental cost-effectiveness ratios are generated through a chain of assumptions that cannot be subjected to direct empirical evaluation. The outputs may be internally consistent, but consistency is not evidence. Scientific claims require measurement, observation, replication and falsification. The reference case offers none of these.

This raises a fundamental question regarding the future of HTA. Can a discipline continue indefinitely when its central constructs fail the standards required for quantitative inference? Perhaps in the short term. Institutions become accustomed to established methods. Journals, agencies, academic programs and consulting organizations develop professional and financial commitments to prevailing frameworks. Yet institutional acceptance does not establish scientific validity. History contains numerous examples where methodological conventions survived for decades before their foundational weaknesses became impossible to ignore.

The alternative is not difficult to identify. Health technology assessment must return to measurement. Every claim for therapy impact must refer to an attribute. That attribute must be identified as either manifest or latent. Manifest attributes require linear ratio measures. Latent attributes require Rasch logit ratio measures. Claims must be formulated as evaluable hypotheses capable of replication, reproduction and falsification in real populations over defined time horizons. Evidence must replace simulation. Measurement must replace arithmetic.

The future of HTA therefore depends upon a choice. One option is to continue defending utilities, QALYs and reference-case simulations despite the absence of demonstrated measurement

foundations. The other is to reconstruct the discipline around the principles that govern every other quantitative science. The issue is no longer whether the reference case can be refined. The issue is whether a framework built upon utility scores and QALYs has any scientific credibility at all. If measurement precedes arithmetic, as it must, then the answer appears increasingly clear; it nver had a future.

THE OBVIOUS SOLUTION: MEASUREMENT BEFORE ARITHMETIC

The most striking feature of the utility framework is that none of its complexity was necessary. The journey from TTO valuations to econometric models, decrement weights, utility algorithms, QALYs and reference-case simulations was undertaken in an attempt to solve a problem that should never have arisen. At every stage increasingly elaborate arithmetic was employed to compensate for the absence of lawful measurement. The result was a framework of impressive mathematical sophistication built upon assumptions that were never demonstrated.

All of this could have been avoided.

The starting point for health technology assessment should never have been preferences, utilities or simulation models. The starting point should have been the attribute. What exactly is the impact of therapy? What attribute is changing? How can that attribute be measured? These are the questions that define every quantitative science. They should also define HTA.

Once the attribute is identified, the measurement problem becomes remarkably simple. There are only two classes of attributes relevant to therapy assessment: manifest attributes and latent attributes.

Manifest attributes are directly observable. Survival time, hospital admissions, emergency department visits, physician consultations, hospital days, laboratory values and medication use are familiar examples. These attributes require linear ratio measures. The standards for their measurement are well established. Once measured, claims regarding therapy impact can be evaluated directly through observation, replication and falsification.

Latent attributes require a different approach. Pain, fatigue, depression, anxiety, physical functioning and need fulfilment cannot be observed directly. They must be inferred from observable responses. For these attributes there is only one acceptable measurement framework: Rasch measurement theory. The outcome is a Rasch logit ratio measure representing possession of the latent attribute. As with manifest attributes, claims can then be subjected to empirical evaluation, replication and falsification.

The implications are profound. Once this distinction is recognized, the elaborate machinery of utility construction becomes unnecessary. There is no need for preference elicitation exercises. There is no need for econometric decomposition of hypothetical health-state valuations. There is no need for decrement weights. There is no need for utility algorithms. There is no need for QALYs. Most importantly, there is no need for reference-case simulations that generate imaginary lifetime claims which cannot be empirically evaluated.

The tragedy of modern HTA is that for over 40 years it chose arithmetic before measurement. Rather than establishing lawful measures and then considering appropriate arithmetic operations, it attempted to create measurement through arithmetic. The result was measurement inversion. Preference scores became utilities. Utilities became QALYs. QALYs became simulation outputs. At no stage was the underlying requirement for lawful measurement satisfied.

The solution is not methodological refinement. It is a return to first principles. Identify the attribute. Determine whether it is manifest or latent. Construct the appropriate ratio measure. Formulate evaluable claims. Test those claims in real populations over defined time horizons. This is the approach adopted throughout the physical sciences and increasingly throughout the social sciences. There is no reason for HTA to operate under different standards.

The lesson is therefore simple. Measurement precedes arithmetic. Once that principle is accepted, the apparent complexity of therapy impact assessment disappears. There are only two measurement paths: linear ratio measures for manifest attributes and Rasch logit ratio measures for latent attributes. Everything else is a diversion. The future of HTA lies not in increasingly sophisticated arithmetic but in the rediscovery of measurement.

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] Wootton D. The Invention of Science: A New History of the Scientific Revolution, New York: HarperCollins, 2015

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

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

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

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