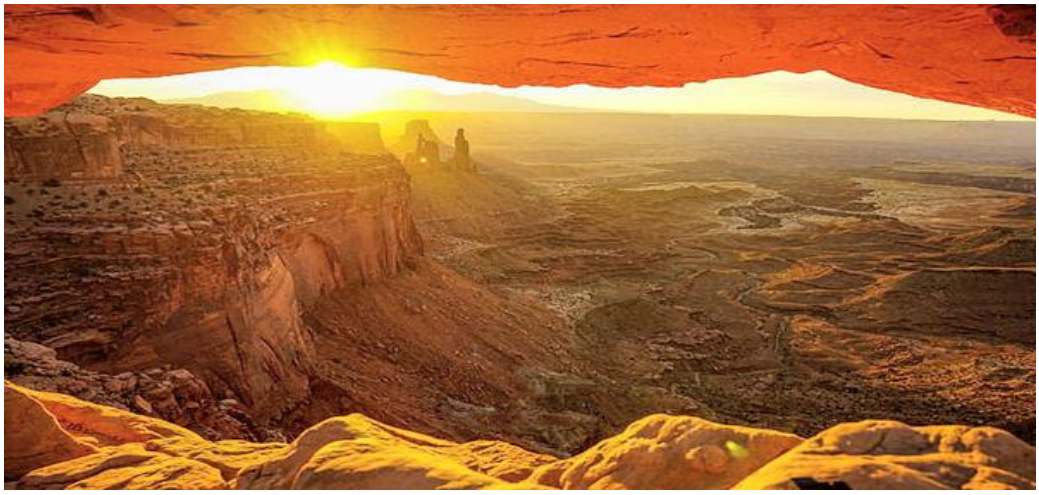
MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION



REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

UNITED STATES: THE SF-36 — FROM DESCRIPTION TO THE SF-6D AND FALSE MEASUREMENT

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

LOGIT WORKING PAPER No 44 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

ABSTRACT

Health technology assessment relies extensively on numerical outcomes derived from patient-reported instruments, yet little attention has been given to how such numbers acquire quantitative authority. This paper examines the SF-36 family of instruments as a critical case study in the epistemic transformation of descriptive scores into apparent measures. Rather than evaluating the instrument's clinical usefulness or psychometric performance, the analysis focuses on the knowledge environment within which SF-36 outputs are interpreted and manipulated as quantities.

The paper argues that the SF-36 occupies a pivotal position in the historical development of modern health technology assessment. Designed as a multidimensional health status profile, the instrument was never intended to measure health as a single quantitative attribute. Its item responses are ordinal, its domains heterogeneous, and its scoring procedures descriptive rather than representational. Nevertheless, over time, its numerical outputs came to be treated as magnitudes. Domain scores were averaged, differences interpreted as effects, and changes over time regarded as improvements or deteriorations, despite the absence of measurement units or invariant structure.

This transformation did not occur through explicit theoretical justification. It emerged through routine use. Repetition within clinical trials, normalization in academic publishing, and transmission through education and analytic software gradually conferred numerical legitimacy on scores that had not been shown to satisfy the axioms of measurement. Statistical modeling, particularly through the development of physical and mental component summary scores, further entrenched this authority by substituting covariance structure for empirical magnitude.

The paper traces how this normalization of scoring prepared the ground for the subsequent conversion of SF-36 data into utilities through the SF-6D. This step is shown not to correct the absence of measurement, but to extend it. Preference weighting replaces description with valuation, yet valuation expresses attitudes toward health states rather than quantities possessed by individuals. The resulting utilities therefore inherit the non-measurement properties of the original scores while enabling their incorporation into cost-utility analysis.

By situating the SF-36 family within its epistemic context, the analysis demonstrates that the core failure is structural rather than technical. Measurement theory does not function as a governing constraint within the knowledge base that sustains instrument use. Arithmetic proceeds without prior establishment of scale type, unidimensionality, or invariance. The paper concludes that the widespread acceptance of SF-36-derived quantities represents an early and decisive step in the emergence of numerical storytelling in health technology assessment, one that made later reliance on utilities and QALYs appear both natural and unproblematic. Recognizing this trajectory is essential for any attempt to recognize representational measurement,

I. INTRODUCTION: THE NORMALIZATION OF SCORING AS MEASUREMENT

Health technology assessment did not begin with utilities, quality-adjusted life years, or reference-case modeling. Its epistemic foundations were laid earlier and more quietly, through the gradual normalization of questionnaire scores as if they were quantitative measures. Long before economic evaluation became institutionalized, health outcomes research had already accepted a critical assumption: that numbers derived from patient-reported instruments could be treated as magnitudes. Once this assumption took hold, the later emergence of utilities and QALYs appeared natural rather than radical.

The SF-36 occupies a pivotal position in this history [i] [ii] [iii] [iv] .Developed as a health status profile, the instrument was intended to describe multiple dimensions of functioning rather than to measure a single latent attribute. Its early success lay in its apparent comprehensiveness and clinical relevance. By organizing patient responses into domains such as physical functioning, pain, vitality, and mental health, the SF-36 offered a structured way to summarize patient experience. It did not claim to measure health as a quantity. It provided descriptive information.

Yet description alone does not explain the influence the SF-36 would come to exert. Over time, its numerical outputs began to circulate as if they were measures. Domain scores were averaged, differences were compared, and changes over time were interpreted as effects. These operations were rarely presented as theoretical claims. They were treated as routine analytic steps. Through repetition, numerical manipulation became normalized, and normalization substituted for justification.

This transformation did not occur through explicit argument. No foundational debate established that ordinal questionnaire responses could support arithmetic operations. Instead, acceptance emerged pragmatically. Scores appeared numeric. Statistical software accepted them. Journals published analyses. Reviewers rarely objected. Each use reinforced the next. What began as descriptive scoring gradually acquired the appearance of measurement. No one mentioned the representational theory of measurement formalized in the early 1970s or the Rasch 1960 model with the rules for transforming observations and scores to interval measurement [v] [vi] [vii] .

The distinction between scores and measures is not semantic. Scores represent counts or aggregates of ordered responses. Measures represent quantities possessing defined units, invariance, and permissible arithmetic which was made clear in Stevens' seminal 1946 contribution [viii] . The former require no theory of magnitude; the latter do. Conflating the two permits calculation but not quantification. Once conflated, however, the difference becomes difficult to recover. Arithmetic performed often enough comes to be perceived as evidence of legitimacy rather than as a practice requiring justification.

The SF-36 played a decisive role in this epistemic shift because it straddled the boundary between description and quantification. Its multi-domain structure invited aggregation, and its numerical formatting encouraged comparison. Although its developers emphasized health profiles rather than summary indices, subsequent practice moved steadily toward numerical condensation. The creation of domain scores, followed by composite physical and mental

summary measures, accelerated this drift. Statistical modeling was introduced to support aggregation, and modeling was mistaken for measurement.

This process reflects a broader pattern in applied science: when numerical representation precedes theoretical constraint, method replaces meaning. Analysts learn how to compute before they learn what computation presupposes. Over time, the appearance of rigor displaces the requirements of representation. What matters is not whether numbers correspond to empirical structure, but whether they behave conveniently within analytic frameworks.

By the time health technology assessment emerged as a formal evaluative discipline, this epistemic groundwork was already in place. Analysts were accustomed to treating questionnaire-derived numbers as quantities. The leap from summated scores to utilities therefore appeared incremental rather than categorical. Valuation did not introduce quantification; it merely repackaged existing numerical assumptions under a new label. The later construction of QALYs would have been inconceivable without this earlier acceptance of scoring as measurement.

This paper argues that the SF-36 represents the hinge point in this historical drift. It did not create the QALY, but it made the QALY plausible. By normalizing arithmetic on non-measures, it trained an entire research culture to accept numerical storytelling as quantitative science. Once that training was complete, the transition from scores to utilities required little resistance.

The purpose of this paper is therefore not to critique the SF-36 as a flawed instrument, nor to question the intentions of its developers. It is to examine the epistemic environment that transformed a descriptive profile into a source of numerical authority. The central question is not whether the SF-36 is useful, widely adopted, or clinically intuitive. It is whether the knowledge base that governs its use recognizes the conditions required for measurement.

By treating the SF-36 as an epistemic object rather than as a technical artifact, the analysis shifts attention from psychometric performance to conceptual legitimacy. The sections that follow trace how ordinal responses were converted into scores, how scores came to be treated as quantities, and how statistical modeling substituted for representational justification. A subsequent section examines the SF-6D as the formal conversion of scores into utilities, revealing the continuity rather than the rupture between descriptive scoring and economic valuation.

Through this structure, the paper seeks to make visible what routine practice has rendered invisible: that the path from questionnaire to QALY did not begin with economics. It began with the unexamined assumption that numbers derived from questionnaires must already be measures.

II. THE SF-36 KNOWLEDGE BASE

To understand how the SF-36 came to function as a source of numerical authority, it is necessary to define the knowledge base within which its outputs are interpreted. That knowledge base cannot be reduced to the original development papers or to the stated intentions of the instrument's authors. Once released into applied research, the SF-36 became embedded within a

distributed epistemic environment that extended far beyond its origins. Its authority emerged not from design alone, but from the collective practices that treated its scores as quantities.

The foundational SF-36 literature introduced the instrument as a health status profile. Its purpose was descriptive: to capture multiple domains of functioning and well-being using standardized questionnaire items. The instrument did not claim to measure a single latent construct, nor did it assert that its scores possessed quantitative properties. This distinction is important. The early framing emphasized breadth, comparability, and interpretability, not measurement in the representational sense.

However, the subsequent life of the SF-36 unfolded largely outside this original framing. As the instrument diffused into clinical trials, outcomes research, and population studies, its numerical outputs began to circulate independently of their descriptive origins. Domain scores were routinely reported as means. Differences were interpreted as effects. Changes over time were treated as improvements or deteriorations. These practices were rarely accompanied by discussion of what kind of numbers were being manipulated. The assumption that scores could function as quantities became implicit.

The applied research literature constitutes the largest component of the SF-36 knowledge base. Thousands of publications employ SF-36 domain scores as dependent variables, predictors, or comparative endpoints. Statistical significance testing, regression modeling, and longitudinal analysis are performed as if the scores possessed invariant units. The sheer volume of such applications confers authority through repetition. When numerical operations appear ubiquitously across high-quality journals, their legitimacy becomes taken for granted.

Editorial and peer-review practices reinforce this normalization. Manuscripts reporting SF-36 scores are seldom challenged on the grounds of scale type or measurement validity. Review typically focuses on sample size, statistical technique, and interpretation of effects, not on whether arithmetic operations are permissible. In this way, methodological scrutiny is displaced from representation to computation. What matters is how numbers are analyzed, not whether they are measures.

Educational transmission plays a central role in stabilizing this knowledge base. In training programs for clinical researchers, epidemiologists, and health economists, the SF-36 is introduced as a standard outcome instrument. Students learn how to score it, how to analyze it, and how to interpret changes. They are not taught to ask whether summated ordinal responses constitute quantities. By the time these analysts enter professional practice, the numerical status of SF-36 scores has already been internalized as unquestioned fact.

The knowledge base is further reinforced through analytic infrastructure. Statistical software, scoring manuals, and published algorithms operationalize the instrument in ways that obscure conceptual assumptions. Once scoring procedures are encoded, users interact with outputs as finished numbers rather than as constructed artifacts. The act of computation becomes separated from the epistemic conditions that would authorize it.

A particularly influential component of the SF-36 knowledge base is the development and dissemination of summary measures. The introduction of physical and mental component scores marked a turning point. These composites were presented as simplified representations of overall health dimensions, derived through statistical modeling. Their availability encouraged further numerical condensation, reinforcing the perception that complex health experiences could be meaningfully reduced to single indices.

Crucially, these composite scores gained authority not through demonstration of measurement properties, which would have failed completely given the axioms of representational measurement, but through institutional uptake. They were incorporated into trials, comparative studies, and meta-analyses. Once embedded in routine practice, their numerical legitimacy was no longer examined. The modeling techniques used to generate them were treated as substitutes for representational justification.

The SF-36 knowledge base thus exhibits a characteristic structure. It is not unified by explicit theoretical agreement about measurement. There is no shared articulation of scale type, unidimensionality, or invariance. Instead, unity arises through coordinated practice. Numbers are used in the same way across studies, institutions, and training environments. This convergence creates the appearance of epistemic stability even in the absence of foundational constraint.

Importantly, this structure does not reflect error or misunderstanding. It reflects non-possession. The principles that determine when numbers can represent magnitudes do not function as governing authorities within the system. Where such principles are absent, they cannot constrain practice. Arithmetic proceeds not because rules are violated, but because the rules are not part of the disciplinary grammar.

Defining the SF-36 knowledge base in this way clarifies the task of analysis. The objective is not to assess whether particular studies misuse the instrument, nor to reinterpret developer intent. It is to interrogate whether the epistemic environment that authorizes numerical use contains the axioms required for measurement. Only by addressing the knowledge base as a whole can the transformation of descriptive scores into quantitative surrogates be properly understood.

The following section therefore turns to the architecture of the SF-36 itself, examining how ordinal item responses are aggregated into domain scores and how scoring practices facilitate the transition from description to numerical authority.

III. THE ARCHITECTURE OF THE SF-36: ORDINAL ITEMS AND SUMMATED SCORES

The epistemic authority of the SF-36 is grounded not in a claim to measurement, but in the numerical form of its outputs. To understand how descriptive responses came to be treated as quantities, it is therefore necessary to examine the internal architecture of the instrument itself. This examination does not evaluate psychometric adequacy or clinical relevance. It addresses a prior and more fundamental question: what kind of numbers does the SF-36 actually produce?

At the item level, the SF-36 consists of questions with ordered response categories. Respondents are asked to indicate frequency, intensity, or limitation using options such as “all of the time,” “most of the time,” or “none of the time.” These categories establish rank order but do not specify distance. The difference between adjacent response options is not defined, constant, or empirically verified. As such, individual item responses are ordinal. They permit ordering, but they do not support arithmetic operations.

This property is not a defect. Ordinal response formats are appropriate for capturing subjective judgments. The difficulty arises not at the level of data collection, but at the level of interpretation. Ordinal responses do not contain units. They cannot be added, averaged, or differenced in a meaningful way unless transformed through a model that establishes invariant measurement. The SF-36 provides no such transformation.

Despite this, item responses are routinely combined into domain scores. Each domain aggregates responses from multiple items using simple summation or linear transformation. The resulting scores are then rescaled, often to a 0–100 range, creating the appearance of continuous measurement. Yet rescaling does not create units. It merely changes numerical labels. An ordinal structure remains ordinal regardless of how it is expressed.

The act of summation is therefore epistemically decisive; so decisive that it fails the axioms if representational measurement. Summation presupposes additivity. Additivity presupposes measurement. In the absence of demonstrated invariance and equal units, summation is not an analytic operation but a numerical convenience. The SF-36 architecture performs this operation implicitly, without articulating or justifying the assumptions it requires.

The resulting domain scores are often interpreted as magnitudes of health status. Higher scores are taken to indicate better functioning, lower scores worse functioning. Differences between scores are interpreted as changes, and comparisons across groups are treated as meaningful contrasts. Yet none of these interpretations follows logically from ordinal data. Ordering alone cannot support claims about amount.

This slippage from order to magnitude is facilitated by the structure of the domains themselves. Each domain combines heterogeneous content under a single label, such as physical functioning or mental health. These domains do not represent latent constructs defined by a common underlying attribute. They are thematic groupings. Items within a domain may differ substantially in difficulty, relevance, and conceptual meaning. Without unidimensionality, there can be no single quantity to measure.

The numerical coherence of domain scores is therefore assumed rather than established. Their stability is inferred from internal consistency statistics or factor loadings, yet such statistics assess correlation, not magnitude. Reliability coefficients describe repeatability of rank order, not existence of units. Factor analysis identifies patterns of covariance, not quantitative structure. None of these techniques can transform ordinal responses into measures.

Nevertheless, the appearance of numerical sophistication exerts powerful influence. Domain scores are expressed as numbers. They vary smoothly. They can be plotted, modeled, and

summarized. This visual and analytic tractability encourages treatment as quantities even when the theoretical conditions for quantification are absent. Numerical behavior substitutes for representational justification.

This substitution is reinforced by norm-based scoring. Domain scores are often standardized relative to population means and variances. Such normalization facilitates comparison across studies but does not create measurement units. Standard deviations are statistical properties, not measurement units. Norm referencing describes position within a distribution, not magnitude of an attribute.

Through these procedures, the SF-36 produces what might be termed numerical surrogates: values that behave like measures within analytic systems without possessing the properties that define measurement. They can be manipulated, but their manipulation does not correspond to lawful transformation of an empirical attribute.

Crucially, this architecture does not require explicit endorsement of measurement claims. The instrument never asserts that it measures health in a representational sense. The transformation occurs downstream, through use. Analysts treat scores as quantities because analytic conventions encourage such treatment. Over time, the distinction between scoring and measurement erodes.

The SF-36 architecture thus provides the material foundation for epistemic drift. Ordinal responses are aggregated, rescaled, and normalized until their origins are obscured. What remains are numbers that invite arithmetic. Once arithmetic becomes routine, the absence of measurement axioms becomes invisible.

This section establishes a critical point for the analysis that follows: the SF-36 does not fail measurement because it is poorly designed. It fails because its architecture does not and cannot establish quantity. The next section examines how this limitation was not merely tolerated but amplified through the creation of composite summary measures, marking the decisive transition from descriptive scoring to numerical authority.

IV. THE PCS/MCS TURN: STATISTICAL MODELING AS SURROGATE MEASUREMENT

The transformation of the SF-36 from a descriptive profile into an apparent quantitative instrument reached its critical moment with the introduction of the Physical Component Summary (PCS) and Mental Component Summary (MCS) scores. These composites were presented as a methodological advance, offering simplified indices that captured overall physical and mental health. In practice, they marked a decisive epistemic shift: statistical modeling was substituted for measurement.

The rationale for the summary scores was pragmatic. Multi-domain profiles were viewed as cumbersome for analysis and communication. Researchers sought parsimonious indicators that could be used in regression models, comparative studies, and longitudinal analyses. The PCS and MCS promised exactly that. By condensing multiple domains into two numbers, they appeared to transform descriptive complexity into analytic clarity.

The method used to generate these scores relied on factor analysis and weighted aggregation. Domain scores were combined using coefficients derived from population covariance structures. The resulting values were then standardized relative to population norms. This procedure produced numbers that were smooth, continuous, and statistically tractable. Yet none of these features establishes measurement.

Factor analysis identifies patterns of correlation. It does not identify magnitude. Loadings reflect shared variance, not units of quantity. A factor score is not an amount of an attribute; it is a weighted position within a statistical structure. Treating such scores as quantities confuses statistical association with empirical measurement.

This confusion was consequential. The PCS and MCS were increasingly interpreted as measures of physical and mental health. Differences were interpreted as changes in health status. Effect sizes were calculated. Regression coefficients were reported. The numerical behavior of the scores encouraged the belief that they represented magnitudes. Yet the underlying data remained ordinal, the domains remained heterogeneous, and no invariant transformation model had been introduced.

The creation of summary scores therefore did not resolve the measurement problem inherent in the SF-36 architecture. It concealed it. By introducing sophisticated statistical machinery, the need for representational justification was displaced. Modeling came to stand in for measurement.

This substitution reflects a broader epistemic error: the assumption that statistical transformation can generate quantity. It cannot. Statistics operate on numbers; measurement determines what numbers represent. Without prior establishment of scale type and invariance, statistical operations merely rearrange symbols. They cannot conjure units where none exist.

The PCS/MCS framework also introduced additional incoherence. The weighting schemes are population-dependent, meaning that the same individual response pattern can yield different summary scores depending on the reference population used. This violates invariance, a core requirement of measurement. A quantity must not change when the population changes. Yet PCS and MCS values are explicitly norm-referenced. Their meaning is relational, not intrinsic.

Moreover, the negative weighting of certain domains in the computation of summary scores produces counterintuitive results, whereby improvements in one domain can reduce the composite score. Such behavior is not anomalous within statistical models, but it is impossible within measurement systems. Quantities cannot decrease when the underlying attribute increases unless measurement has failed.

These features did not provoke epistemic alarm. Instead, they were absorbed as technical nuances. The authority of the scores rested not on their coherence as measures, but on their usefulness within analytic workflows. Once embedded in software, guidelines, and published literature, the summary scores became routine objects of inference.

The PCS/MCS turn therefore represents the moment at which numerical authority became detached from empirical representation. Scores were no longer merely descriptive aggregates; they became surrogates for latent attributes. Health was inferred not from possession of an attribute, but from position within a statistical construct.

This shift had far-reaching consequences. By accepting composite scores as quantities, the research community became habituated to the idea that latent constructs could be quantified through statistical modeling alone. The distinction between measurement and estimation blurred. What mattered was not whether an attribute had been measured, but whether a number could be generated.

This epistemic accommodation prepared the ground for the next transformation: the conversion of questionnaire scores into utilities. Once it was accepted that statistical models could generate quantities from ordinal data, the leap to preference-weighted utilities appeared modest. Valuation seemed merely an extension of scoring, rather than a categorical shift.

The PCS/MCS framework thus occupies a pivotal position in the genealogy of health technology assessment. It did not introduce utilities or QALYs, but it normalized the belief that numbers derived from questionnaires could legitimately function as measures. By the time economic evaluation formalized this belief, its foundations were already secure.

The following section examines this transition explicitly through the development of the SF-6D. By tracing how SF-36 scores were converted into utilities, the analysis demonstrates that valuation did not correct the epistemic failure introduced by scoring. It merely repackaged it under a different name.

V. FROM SCORES TO UTILITIES: THE SF-6D CONVERSION

The development of the SF-6D represents the formal transition from descriptive scoring to economic valuation [ix]. Where the SF-36 produced domain scores and statistical composites, the SF-6D sought to generate utilities suitable for cost-utility analysis. This step is often portrayed as a methodological innovation that enabled the use of existing health status data in economic evaluation. From an epistemic perspective, however, it represents something more consequential: the conversion of non-measures into quantities by decree.

The SF-6D was constructed by selecting a subset of dimensions from the SF-36, defining a reduced health state classification system, and assigning preference weights derived from population valuation exercises. The resulting algorithm produces numerical values anchored at full health and death and expressed on a scale conventionally treated as interval or ratio. These values are then multiplied by time to generate QALYs.

What distinguishes this process is not its technical sophistication but its foundational assumption. The SF-6D presumes that valuation can substitute for measurement. It assumes that attaching preferences to descriptive states creates magnitude where none previously existed. Yet valuation expresses order and intensity of preference, not quantity of an attribute. Preferences do not measure health; they reflect judgments about health states made by observers.

This distinction is not semantic. Measurement concerns properties of persons. Valuation concerns attitudes toward hypothetical descriptions. When preference weights are applied to SF-36-derived states, the numerical output reflects how much a population prefers one description to another, not how much health an individual possesses. The SF-6D therefore does not transform scores into measures. It replaces description with valuation.

From the standpoint of representational measurement theory, this substitution is illegitimate. No invariant transformation model is introduced. The ordinal nature of the underlying responses remains unchanged. Unidimensionality is not established. No empirical attribute is identified whose magnitude is preserved under numerical operations. The resulting utilities behave numerically, but their behavior is imposed rather than derived.

The SF-6D thus inherits every limitation of the SF-36 architecture while introducing additional contradictions. Its health state classification remains multiattribute. Its valuation model aggregates heterogeneous dimensions into a single number without establishing a common unit. Its allowance of negative values contradicts ratio-scale requirements. Yet these contradictions do not inhibit use. They are absorbed into analytic convention.

Crucially, the SF-6D does not represent a conceptual rupture from the SF-36. It represents continuity. The epistemic move that enables SF-6D is the same one that enabled PCS and MCS scores: the belief that numerical manipulation confers quantitative meaning. Once that belief is in place, valuation appears as merely another transformation step.

This continuity explains why SF-6D utilities exhibit structural invariance with other preference-based instruments such as EQ-5D, HUI, and AQoL. Despite differences in descriptive systems and valuation protocols, all generate utilities that behave in the same way because they share the same epistemic foundation. None measure an attribute. All assign numbers to descriptions and treat the result as quantity.

The transition from SF-36 to SF-6D therefore completes the epistemic drift traced throughout this paper. What began as descriptive profiling evolved into summated scoring, then into statistical composites, and finally into utilities. At no stage was measurement established. Each step merely extended numerical authority while moving further from representational grounding.

This genealogy clarifies why later debates over valuation methods, population preferences, or mapping algorithms cannot resolve the underlying problem. Such debates occur entirely within a framework that assumes measurement has already occurred. The SF-6D demonstrates that this assumption is false. Valuation cannot repair the absence of quantity; it can only disguise it.

The SF-6D is thus not an isolated methodological innovation. It is the logical endpoint of a process in which scoring gradually came to be mistaken for measurement. By converting SF-36 data into utilities, it provided the missing link between health status questionnaires and cost-utility analysis, making the QALY appear empirically grounded when it was not.

The final section draws these threads together. It considers the implications of this epistemic trajectory for health technology assessment and argues that the problem exposed by the SF-36

family is not technical, but structural. The issue is not how numbers are produced, but what they are taken to represent.

VI. CONCLUSIONS: FROM DESCRIPTIVE SCORES TO NUMERICAL STORYTELLING

This paper has traced the epistemic trajectory through which the SF-36 family of instruments came to occupy a position of numerical authority within health technology assessment. The analysis has not questioned the clinical usefulness of descriptive health profiles, nor the intentions of their developers. Instead, it has examined how numbers derived from questionnaires acquired quantitative meaning without ever satisfying the conditions required for measurement.

The central finding is that the SF-36 did not fail as a measurement instrument because it was poorly designed. It was never designed to measure. Its purpose was descriptive. The failure occurred later, through use. Ordinal responses were aggregated into scores, scores were condensed into statistical composites, and composites were converted into utilities. At no stage was the representational problem addressed. Measurement was presumed rather than established.

This process reveals an epistemic drift rather than a methodological error. Each step appeared incremental. Summation followed description. Statistical modeling followed summation. Valuation followed modeling. At no point did the transition appear radical enough to provoke foundational scrutiny. Yet collectively these steps produced a profound inversion: arithmetic became primary, while measurement receded from view.

The introduction of PCS and MCS summary scores marked a decisive moment in this drift. By treating factor-derived composites as quantities, the research community accepted statistical structure as a surrogate for empirical magnitude. Once that substitution was normalized, the conversion of questionnaire outputs into utilities required no conceptual leap. Valuation appeared to complete what scoring had already begun.

The SF-6D demonstrates the consequence of this normalization. It does not measure health. It assigns preference values to descriptive states and treats the result as quantity. In doing so, it inherits every limitation of the SF-36 while extending its numerical reach into economic evaluation. The apparent continuity between SF-36 and SF-6D is therefore not evidence of conceptual coherence. It is evidence of structural invariance in epistemic failure.

This invariance links the SF-36 family directly to the broader architecture of health technology assessment. Preference-based instruments such as EQ-5D, HUI, and AQoL differ in surface features but share the same foundational assumption: that valuation can substitute for measurement. The present analysis shows that this assumption did not originate in economics. It was learned earlier, through decades of scoring practice that accustomed researchers to treating numbers as quantities simply because they could be computed.

The implications for HTA are substantial. When instruments that do not measure are nevertheless treated as quantitative, all downstream analysis becomes epistemically unstable. Cost-utility models inherit numbers that lack units. QALYs are constructed through arithmetic

that has no lawful foundation. Precision is simulated through calculation, not secured through measurement.

This conclusion does not call for refinement of scoring algorithms, improved valuation surveys, or alternative preference elicitation techniques. None of these can supply what is missing. Measurement cannot be added after the fact. It must precede arithmetic. Where unidimensionality, invariance, and scale-type coherence are absent, no amount of statistical or economic sophistication can restore quantitative meaning.

The contribution of this paper lies in making explicit a process that has long remained implicit. By situating the SF-36 within its epistemic environment, the analysis demonstrates how numerical authority is socially constructed and institutionally reinforced. The problem is not ignorance, nor technical incompetence. It is non-possession of measurement theory as a governing constraint.

Recognizing this condition is a necessary first step toward reform. If health technology assessment is to claim scientific legitimacy, it must recover the distinction between description and measurement, between scoring and quantity, between valuation and possession. Without that recovery, numerical storytelling will continue to masquerade as quantitative science.

The SF-36 did not create this problem. It revealed it. And in doing so, it provides a clear starting point for rebuilding HTA on foundations that can support the arithmetic it so confidently deploys.

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

[i] Ware, J., Sherbourne C. The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual Framework and Item Selection 1992 *Medical Care* 30 (6): 473–483

[ii] Lins L, Carvalho G. SF-36 Total Score as a Single Measure of Health-Related Quality of Life: Scoping Review. *Health and Quality of Life Outcomes* 1 2016; 4 (*SAGE Open Medicine* 4 (October): 2050312116671725)

[iii] Ware J et al. 1993. *SF-36 Health Survey: Manual and Interpretation Guide.* Boston: Health Institute, New England Medical Center

[iv] Maruish M. 2011. *User's Manual for the SF-36v2 Health Survey.* 3rd ed. Lincoln, RI: QualityMetric Incorporated

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

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

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

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

[ix] Brazier J, Roberts J, Deverill**.** The Estimation of a Preference-Based Measure of Health from the SF-36. *J Health Economics* 2022; 21(2): 271–292