MAIMON RESEARCH LLC

ARTIFICIAL INTELLIGENCE LARGE LANGUAGE MODEL INTERROGATION

REPRESENTATIONAL MEASUREMENT FAILURE IN HEALTH TECHNOLGY ASSESSMENT

AUSTRALIA: PSYCHOMETRICS IS NOT MEASUREMENT

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

LOGIT WORKING PAPER No 727 FEBRUARY 2026

www.maimonresearch.com

Tucson AZ

PSYCHOMETRICS IS NOT MEASUREMENT

INTRODUCTION

A defining feature of the knowledge bases maintained not only by international clinical and outcomes organizations, but also by national health systems, HTA agencies, regulatory authorities, academic research centers, and major journals, is their near-universal reliance on psychometric scoring systems to evaluate therapy impact. Across disease areas, patient-reported outcome instruments are constructed, validated, and applied using frameworks that treat summed subjective response scores as quantitative measures. These scores underpin comparative effectiveness claims, cost-effectiveness analyses, and reimbursement decisions. The approach is administratively convenient and institutionally entrenched, yet it rests on a fundamental error: ordinal scores do not constitute measurable quantities.

The intuitive appeal of assigning numbers to subjective responses creates the appearance of quantification. The resulting scores can be averaged, compared, and incorporated into statistical analyses. Over time, repeated use of these scoring systems has conferred legitimacy through familiarity and institutional endorsement. Psychometric validation frameworks reinforce this legitimacy by emphasizing reliability, internal consistency, and construct validity. These criteria create the impression that instruments producing stable and interpretable scores must therefore be measuring underlying attributes. However, stability and interpretability do not establish measurement. They do not demonstrate that the numerical values assigned to observations possess invariant quantitative meaning. Without invariant unit structure and dimensional homogeneity, scores remain ordered classifications rather than quantities.

The distinction between scoring and measurement is not optional. It is defined by the axioms of representational measurement theory, which specify the conditions under which numbers legitimately represent empirical quantities [1] [2]. Measurement requires more than numerical assignment; it requires that numerical relations correspond to empirical relations in a structurally invariant manner. Only under these conditions can arithmetic operations such as addition, multiplication, and ratio comparisons be interpreted meaningfully. Psychometric scores do not satisfy these conditions. They preserve order, but they do not establish quantity. As a consequence, arithmetic operations performed on these scores lack empirical interpretation, even when they are statistically convenient.

This problem has profound implications for therapy evaluation. International clinical organizations play a central role in defining outcome standards, developing instruments, and shaping the evidentiary foundation for health technology assessment. When these organizations adopt scoring systems in place of measurement, the consequences propagate throughout the evaluative architecture of modern medicine. Clinical trials, regulatory submissions, and reimbursement decisions inherit numerical constructs that possess administrative utility but lack measurement validity. This is not the result of negligence or incompetence, but of historical inheritance. Psychometric scoring systems emerged as practical tools for summarizing subjective experience at a time when lawful measurement methods for latent traits were not widely recognized or applied.

The emergence of representational measurement theory and the Rasch model fundamentally alters this landscape [3] [4]. These frameworks provide explicit criteria for distinguishing scores from measures and establish the only scientifically valid methods for quantifying latent attributes. The four sections that follow examine this distinction in detail. They explain why psychometrics cannot produce measurement, why Rasch transformation is necessary for latent trait quantification, and why the future of therapy evaluation depends on the transition from scoring systems to lawful measurement structures.

SECTION I

MEASUREMENT AND THE AXIOMS OF QUANTIFICATION: WHAT MEASUREMENT IS

Measurement is not the assignment of numbers. Measurement is the demonstration that numbers represent quantities. This distinction, simple in statement but profound in consequence, defines the boundary between quantitative science and numerical description. It is the foundation upon which physics, chemistry, engineering, and every quantitative discipline rests. Without measurement, arithmetic operations are devoid of empirical meaning. Numbers may be manipulated, compared, or modeled, but they do not correspond to quantities in the world. They remain symbols, not measures.

STEVENS AND REPRESENTATIONAL MEASUREMENT THEORY

The formal scientific account of measurement is provided by representational measurement theory. This theory establishes the conditions under which numerical assignments preserve the empirical structure of the attribute being measured. Measurement exists when numerical relations correspond to empirical relations. If one object possesses more of an attribute than another, the assigned number must be greater. If the difference between two objects is equivalent to the difference between two others, the numerical differences must also be equivalent. If ratios between quantities exist empirically, those ratios must be preserved numerically. Measurement is therefore a mapping between empirical reality and numerical structure. It is not an arbitrary assignment but a lawful correspondence.

This principle was recognized long before modern statistical methods emerged [5]. Galileo's observation that the laws of nature are written in the language of mathematics presupposed that quantities exist independently of their numerical representation. Newton's mechanics depended upon measurable quantities such as time, distance, and mass, each possessing invariant unit structure. Without invariant units, Newton's equations would have no empirical meaning. The success of physical science did not arise from the manipulation of numbers, but from the establishment of measurement systems in which numerical operations corresponded to empirical relations.

Representational measurement theory formalized these principles. Stevens' well-known typology distinguished nominal, ordinal, interval, and ratio scales, identifying the operations admissible at each level. Ordinal scales preserve order but do not support arithmetic operations beyond comparison. Interval scales preserve equal differences but lack a true zero. Ratio scales possess a true zero and invariant units, allowing multiplication and division. This classification did not create measurement; it clarified the conditions under which measurement exists. Arithmetic operations are admissible only when the scale structure supports them.

These conditions are not arbitrary conventions. They arise from the structure of empirical attributes themselves. A ratio scale exists when equal units represent equal quantities and when the absence of the attribute corresponds to zero. Length satisfies these conditions. Time satisfies these conditions. Mass satisfies these conditions. Arithmetic operations performed on such quantities

preserve empirical relations. Doubling a length corresponds to twice the empirical distance. Halving a time corresponds to half the empirical duration. Arithmetic reflects empirical reality because measurement has been established.

MEASUREMENT PRECEDES ARITHMETIC

The central principle that follows from representational measurement theory is that measurement precedes arithmetic. Arithmetic operations do not create measurement. They depend upon measurement. Multiplication, division, addition, and subtraction have empirical meaning only when applied to quantities possessing lawful scale structure. When arithmetic is applied to numbers that do not represent quantities, the operations produce numerical outputs without empirical interpretation. The results may appear precise, but precision without measurement is an illusion. It reflects internal numerical coherence, not empirical correspondence.

This principle is universal. It applies equally to manifest and latent attributes. Manifest attributes, such as survival time or resource utilization, can be directly observed and measured using linear ratio scales. Latent attributes, such as pain severity or functional impairment, cannot be directly observed. Their measurement requires transformation of observable indicators into invariant quantities; the Rasch transformation. In both cases, the requirement is the same: numerical assignments must preserve empirical structure. Without this preservation, measurement does not exist.

Measurement therefore imposes strict requirements. The attribute must be unidimensional. Measurement units must be invariant across observations. The numerical representation must preserve empirical relations. These conditions define the existence of quantity. They are not methodological preferences. They are logical necessities. Without unidimensionality, numerical assignments cannot represent a single attribute. Without invariant units, differences between numerical values cannot be interpreted consistently. Without preservation of empirical relations, numerical assignments are arbitrary.

QUANTITY NOT ORDER

These principles distinguish measurement from scoring. Scoring systems assign numbers according to predefined rules. These rules may produce internally consistent numerical outputs. They may correlate with observable phenomena. They may support statistical analysis. But unless the numerical assignments preserve empirical structure, they do not constitute measurement. They produce scores, not measures. Scores describe order. Measures represent quantity. This distinction is fundamental.

The consequences of violating measurement principles are not immediately apparent. Numerical outputs produced by scoring systems possess the appearance of precision. They may be expressed to multiple decimal places. They may be incorporated into complex models. They may be used to generate ratios, projections, and forecasts. But without measurement, these operations lack empirical grounding. They produce numerical artifacts, not quantitative knowledge. The appearance of quantification substitutes for quantification itself.

Scientific progress depends upon measurement because measurement enables falsification [6]. Quantitative claims can be tested against empirical observations only when the quantities being compared possess invariant structure. If a therapy reduces pain by a measurable amount, that reduction must correspond to a difference in measured quantity. If numerical differences do not represent quantities, empirical testing becomes impossible. Claims cannot be falsified because they do not refer to measurable attributes. Numerical outputs exist only within the scoring system that produced them.

SCIENCE AND PSEUDOSCIENCE

Measurement therefore defines the boundary between empirical science and numerical description. Disciplines that establish measurement systems can develop cumulative knowledge. Quantitative claims can be replicated, compared, and refined. Errors can be identified and corrected. Knowledge evolves through empirical testing. Disciplines that lack measurement systems cannot follow this trajectory. Numerical outputs may proliferate, but they cannot be empirically validated. Knowledge becomes detached from measurement.

Representational measurement theory resolved these issues by identifying the formal conditions under which measurement exists. These conditions apply universally. They do not depend upon disciplinary tradition or methodological convenience. They reflect the logical requirements of quantification itself. Measurement cannot be established by statistical analysis alone. Statistical methods operate on numerical inputs. They cannot create measurement where measurement does not exist. Statistical sophistication cannot substitute for measurement validity.

The recognition that measurement precedes arithmetic has profound implications. It means that numerical operations must be restricted to quantities possessing lawful scale structure. It means that composite indices lacking invariant units cannot support multiplication or division. It means that summation of ordinal categories does not produce quantity. It means that arithmetic operations performed on non-measures do not yield quantitative results. These conclusions follow directly from the axioms of representational measurement.

MEASUREMENT AND QUANTITATIVE SCIENCE

The importance of this principle cannot be overstated. Measurement is the foundation of quantitative science [7]. Without measurement, numerical analysis becomes symbolic manipulation detached from empirical reality. Numbers may be produced, but they do not represent quantities. Models may be constructed, but they do not quantify attributes. Claims may be expressed numerically, but they do not constitute measurement.

The distinction between measurement and scoring defines the central problem addressed in this paper. Psychometric systems assign numbers to observations according to predefined rules. These assignments produce scores. Whether those scores constitute measures depends upon whether they satisfy representational measurement axioms. This is not a matter of statistical reliability or predictive validity. It is a matter of measurement existence. Either the numerical assignments preserve empirical structure or they do not.

The subsequent sections will demonstrate that psychometric scoring systems do not satisfy these requirements. They produce internally consistent numerical outputs but do not establish invariant unit structure. Their numerical assignments describe order but do not represent quantity. As a result, arithmetic operations performed on psychometric scores lack empirical meaning. The appearance of measurement substitutes for measurement itself.

The resolution to this problem was provided by Rasch measurement, which establishes invariant unit structure for latent attributes through conjoint simultaneous measurement. Rasch transformation produces logit ratio scales satisfying representational measurement axioms [8] . These scales constitute lawful measures. They differ fundamentally from psychometric scores. They represent quantities rather than numerical rankings.

The distinction between scoring and measurement therefore defines the boundary between psychometrics and measurement science. Psychometrics produces scores. Measurement science produces measures. Only measurement supports arithmetic operations with empirical meaning. Only measurement enables falsification, replication, and cumulative scientific knowledge.

The failure to recognize this distinction has allowed scoring systems to be treated as if they were measurement systems. Numerical outputs derived from ordinal observations have been subjected to arithmetic operations requiring ratio scale properties. This inversion of logical order, performing arithmetic before establishing measurement, lies at the core of the problem addressed in this paper.

Measurement precedes arithmetic. This is not a methodological recommendation. It is the logical foundation of quantitative science.

SECTION II

PSYCHOMETRICS AND THE SCORING FALLACY: WHY PSYCHOMETRIC SCORES ARE NOT MEASURES

Psychometrics did not emerge as a theory of measurement. It emerged as a theory of scoring. Its original purpose was practical: to provide systematic procedures for assigning numerical values to responses on tests and questionnaires. These numerical assignments were intended to summarize patterns of responses, facilitate comparisons, and support statistical analysis. Psychometrics succeeded in this objective. It provided methods for constructing instruments, evaluating internal consistency, assessing test–retest reliability, and examining correlations between scores and external variables. These achievements created the appearance of quantitative rigor. Yet the central question was never resolved: do psychometric scores constitute measurement?

The answer, established by representational measurement theory, is no. Psychometric scores preserve order but do not establish quantity. They are ordinal assignments, not measures. Their numerical values indicate rank position, not magnitude. A higher score indicates more of something, but the difference between scores does not necessarily correspond to an invariant difference in the attribute. The numerical intervals between scores are determined by scoring rules, not by empirical demonstration of equal units.

This distinction is decisive. Measurement requires invariant units. The difference between 10 and 20 must represent the same quantity as the difference between 30 and 40. Without invariant units, numerical differences cannot be interpreted as quantitative differences. Psychometric scoring systems do not establish invariant units. They assign numerical values based on category labels, item responses, or summation rules. These assignments create numerical order, but they do not demonstrate that equal numerical differences represent equal empirical differences.

Consider the summation of Likert-scale responses, the most common psychometric procedure. Respondents select categories such as “strongly disagree,” “disagree,” “neutral,” “agree,” and “strongly agree.” These categories are assigned numerical values, typically 1 through 5. Responses across items are summed to produce a total score. This score is then treated as a quantitative representation of the attribute. Arithmetic operations are performed. Means are calculated. Differences are compared. Ratios may even be constructed. Yet the numerical assignments underlying this process are arbitrary. The difference between categories labeled 1 and 2 is not empirically demonstrated to equal the difference between categories labeled 3 and 4. The numbers preserve order but do not establish quantity.

The summation of ordinal responses does not transform ordinal data into interval or ratio measurement. This principle was established conclusively in measurement theory. Arithmetic operations do not alter the scale properties of the underlying observations. Summing ordinal scores produces another ordinal score. It does not produce measurement. The resulting numerical values remain dependent on scoring conventions rather than empirical structure.

Psychometrics attempts to address this limitation through statistical evaluation. Reliability coefficients such as Cronbach’s alpha assess internal consistency. Factor analysis examines

dimensional structure. Correlation coefficients evaluate relationships with external variables. These statistical procedures provide useful information about score behavior. They assess stability, coherence, and association. But they do not establish measurement. Statistical consistency does not create invariant units. Correlation does not establish quantity. Statistical analysis operates on numerical inputs; it does not determine whether those inputs represent measures.

This distinction is often misunderstood. Reliability is not measurement. A score can be highly reliable yet not represent quantity. Repeated measurements of ordinal rank will produce consistent rankings, but rankings do not constitute measurement. Validity is not measurement. A score can correlate strongly with external criteria yet remain ordinal. Correlation does not create quantity. Statistical properties describe score behavior. They do not establish measurement existence.

SUMMED SCORES ARE NOT INTERVAL MEASURES

Psychometric theory often assumes that summated scores approximate interval measurement. This assumption lacks empirical justification. Approximation is not measurement. Measurement requires demonstration, not approximation. Without invariant unit structure, numerical differences cannot be interpreted quantitatively. Treating psychometric scores as interval or ratio measures constitutes an assumption, not a demonstration.

The consequences of this assumption are profound. Arithmetic operations performed on psychometric scores lack empirical meaning. Differences between scores do not represent differences in quantity. Ratios between scores do not represent ratios of magnitude. Means calculated from scores do not represent average quantities. These operations produce numerical outputs that reflect scoring conventions rather than empirical measurement.

Psychometric scoring systems also fail the requirement of unidimensionality. Measurement requires that numerical assignments represent a single attribute. Psychometric instruments often combine multiple attributes. Questionnaires measuring “quality of life,” “pain,” or “functional status” typically include items addressing different dimensions. Summing responses across these dimensions produces composite scores. These scores do not represent a single quantity. They represent aggregated observations across heterogeneous attributes. Arithmetic operations performed on such composites lack dimensional coherence.

Dimensional homogeneity is a prerequisite for arithmetic. Quantities must represent the same attribute. Adding lengths measured in meters produces meaningful results. Adding lengths and weights does not. Psychometric composite scores combine heterogeneous observations. The resulting numerical values do not represent quantities possessing dimensional homogeneity. Arithmetic operations performed on these scores do not correspond to empirical operations.

Psychometric scoring also lacks invariance. Measurement units must remain constant across individuals and contexts. The meaning of one unit must be the same regardless of who is measured. Psychometric scores do not satisfy this requirement. The same numerical score may represent different empirical conditions depending on the pattern of item responses. Item difficulty and respondent characteristics interact. The numerical assignment depends on the scoring system rather than invariant unit structure.

This lack of invariance prevents meaningful comparison. Differences between scores cannot be interpreted consistently across individuals. Arithmetic operations assume invariant units. Without invariance, arithmetic loses empirical meaning. Psychometric scoring systems do not establish invariant units. They produce context-dependent numerical assignments.

Psychometrics also fails the requirement of a true zero. Ratio measurement requires a zero point representing the absence of the attribute. Psychometric scores typically lack such a point. The lowest score does not represent absence. It represents the lowest possible score within the scoring system. Arithmetic operations such as multiplication and division require ratio scale properties. Psychometric scores do not possess these properties.

PSYCHOMETRIC SCORING SYSTEM PERSISTENCE

The persistence of psychometric scoring systems reflects historical inheritance rather than measurement validation. Psychometrics provided practical tools for summarizing observations. These tools facilitated statistical analysis and instrument development. Their numerical outputs created the appearance of measurement. Over time, this appearance became accepted as measurement itself. The distinction between scoring and measurement was obscured.

This confusion was reinforced by statistical methodology. Statistical procedures operate on numerical inputs regardless of their scale properties. Means, standard deviations, correlations, and regressions can be calculated from ordinal scores. These calculations produce numerical outputs. The existence of numerical outputs creates the impression of quantitative analysis. Yet statistical analysis does not create measurement. It operates on numbers. Whether those numbers represent quantities depends on measurement properties, not statistical procedures.

Psychometric theory did not resolve this problem because it focused on statistical properties rather than measurement axioms. Reliability, validity, and factor structure became central concerns. Measurement existence was assumed rather than demonstrated. The scoring paradigm became institutionalized. Psychometric scores were treated as measures without satisfying representational measurement requirements.

The consequence is the widespread use of numerical scores that do not represent quantities. These scores support statistical analysis, model construction, and numerical comparison. Yet their numerical values reflect scoring conventions rather than empirical measurement. Arithmetic operations performed on these scores produce numerical outputs without quantitative meaning. This condition defines the scoring fallacy: the belief that assigning numbers creates measurement. Psychometric scoring assigns numbers. It does not create invariant unit structure. Without invariant units, numerical assignments do not constitute measurement. They remain ordinal descriptions expressed numerically.

The resolution to this problem requires a measurement model that establishes invariant unit structure for latent attributes. Psychometric scoring does not provide such a model. Its numerical assignments remain dependent on scoring conventions rather than empirical structure. Measurement requires transformation of observations into invariant quantities. This transformation is provided by Rasch measurement.

Rasch measurement differs fundamentally from psychometric scoring. It establishes invariant unit structure through conjoint simultaneous measurement of persons and items. It produces logit ratio scales satisfying representational measurement axioms. Psychometric scoring produces ordinal scores. Rasch measurement produces measures. The distinction between psychometrics and measurement is therefore not methodological but foundational. Psychometrics provides scoring procedures. Measurement science provides measurement systems. Arithmetic operations require measurement systems. Psychometric scores do not satisfy this requirement.

The persistence of psychometric scoring systems reflects institutional continuity rather than measurement validity. Their numerical outputs facilitate statistical analysis and administrative decision-making. Yet their numerical assignments do not represent quantities. Arithmetic operations performed on psychometric scores do not correspond to empirical operations.

Psychometrics is therefore not measurement. It is scoring. The numerical outputs it produces describe order but do not represent quantity. Measurement requires invariant unit structure. Psychometric scoring does not provide this structure. The appearance of measurement substitutes for measurement itself.

SECTION III

RASCH MEASUREMENT: THE ONLY SCIENTIFIC FRAMEWORK FOR MEASURING LATENT TRAITS

The failure of psychometrics to establish measurement does not imply that latent attributes cannot be measured. It implies only that scoring procedures do not constitute measurement. Measurement requires the construction of an invariant quantitative structure in which numerical values correspond to empirical magnitudes. For manifest attributes such as length, mass, and time, this structure emerges through physical operations that establish invariant units. For latent attributes such as pain severity, functional impairment, or need fulfillment, invariant units cannot be established through direct physical operations. They must instead be constructed through a measurement model. The Rasch model provides the only framework capable of achieving this objective.

THE RASCH MODEL

The Rasch model is not a statistical convenience but a measurement theory. Its purpose is to transform ordinal observations into quantitative measures possessing invariant unit structure. This transformation is achieved through conjoint simultaneous measurement of two independent entities: the ability or trait level of persons and the difficulty or severity of items. The model specifies the probability that a person with a given level of the latent trait will endorse or respond affirmatively to an item of given difficulty. This probabilistic relationship defines a quantitative structure in which both persons and items are located on a common continuum.

The fundamental property of Rasch measurement is invariance. The measurement of persons does not depend on the specific items used, and the measurement of items does not depend on the specific persons observed, provided the model fits the data. This invariance establishes unit structure. Differences between logit values represent invariant differences in the latent attribute. The logit scale therefore satisfies the requirements of measurement. Numerical differences correspond to empirical differences. Arithmetic operations performed on logit measures possess empirical meaning.

This property distinguishes Rasch measurement from psychometric scoring. In psychometric scoring, numerical assignments depend on scoring conventions. The meaning of score differences varies with the instrument and population. No invariant unit exists. In Rasch measurement, numerical assignments emerge from the measurement model itself. The unit is defined by the logit, the natural logarithm of the odds ratio relating person ability and item difficulty. This unit remains invariant across contexts. Measurement becomes independent of scoring conventions.

The Rasch model also establishes unidimensionality. Measurement requires that numerical values represent a single attribute. The Rasch model tests this requirement explicitly. If observations do not conform to a single latent continuum, the model will not fit the data. Misfitting items or persons can be identified and removed. This process ensures that the resulting measures represent a single latent attribute. Psychometric scoring does not provide this safeguard. Composite scores often combine multiple attributes without establishing unidimensionality.

The Rasch model also satisfies the requirement of dimensional homogeneity. All logit values represent the same attribute. Arithmetic operations performed on logit measures therefore possess empirical meaning. Differences between logit values represent differences in the latent trait. Comparisons between measures are valid because the unit structure remains invariant.

Another essential property of Rasch measurement is the existence of a true zero point. The logit scale possesses a natural origin corresponding to equal odds of endorsement. This origin allows the construction of ratio comparisons. Because the logit scale is a ratio scale, multiplication and division possess empirical meaning. This property distinguishes Rasch measurement from interval scales, which lack a true zero.

The logit transformation itself is central to this process. The logit is defined as the natural logarithm of the odds of endorsement. Odds represent ratios of probabilities. The logarithmic transformation produces a scale possessing additive properties. Differences in logits correspond to multiplicative differences in odds. This transformation establishes ratio scale structure. Measurement becomes possible because numerical differences correspond to empirical differences in the latent attribute.

The Rasch model also provides a mechanism for testing measurement validity. Model fit statistics assess whether observations conform to the measurement model. If data do not fit the model, measurement has not been achieved. This requirement reflects the principle that measurement must be demonstrated, not assumed. Psychometric scoring does not provide this safeguard. Scores are calculated regardless of whether measurement exists.

INTERPRETING RASCH POSSESSION: LOGITS, SIGNIFICANCE, AND MEASUREMENT VALIDITY

A central misunderstanding in discussions of Rasch measurement concerns the meaning of the logit and the concept of "possession" of a latent trait. In Rasch measurement, the outcome of interest is not a score but a quantified level of possession of a latent attribute. Whether the attribute is pain severity, functional ability, depression, or quality of life, the Rasch model places both persons and items on the same latent continuum expressed in logits. A logit is the natural logarithm of the odds that a person with a given trait level will endorse or succeed on an item of given difficulty. This transformation converts ordinal response probabilities into an interval structure with ratio properties on the logit scale. The result is a quantitative representation of possession.

Possession in Rasch measurement is therefore not metaphorical. It is expressed as a location on a logit ratio scale. Individuals possess more or less of the latent attribute in quantifiable units. Differences between person measures in logits represent invariant differences in the latent trait. Because the logit scale is derived from a ratio of probabilities, it possesses a meaningful zero point (equal odds of endorsement) and supports meaningful comparison. A difference of 1.0 logit between two individuals represents the same quantitative difference anywhere along the scale. This invariance is what distinguishes Rasch measures from psychometric scores.

Group possession can be evaluated in the same manner. Mean logit measures for treatment and control groups can be compared directly. The Rasch framework provides standard errors for each person and group estimate, allowing formal statistical evaluation. Differences in group means can

be tested using t-statistics derived from the difference in logits divided by the pooled standard error. Confidence intervals can be constructed around person or group measures. If the confidence intervals do not overlap, the difference is statistically significant. Thus therapy impact expressed as change in logit possession can be evaluated both quantitatively and inferentially.

Beyond statistical significance, Rasch measurement allows evaluation of substantive significance. Because the logit scale is invariant, the magnitude of change can be interpreted directly. A therapy producing a 0.5 logit improvement reflects a defined shift in the probability structure of item endorsement. This is not a change in a scoring convention but a change in quantified possession of the latent trait.

Equally important, the Rasch model provides an internal mechanism for testing whether measurement has been achieved. Fit statistics assess the degree to which observed responses conform to model expectations. Infit and outfit mean square statistics identify whether persons or items behave consistently with the unidimensional latent structure. If data do not fit the model, the assumption of measurement is rejected. Items may be revised or removed. The model therefore enforces measurement validity rather than assuming it.

This requirement reflects a core scientific principle: measurement must be demonstrated, not assumed. Psychometric scoring lacks this safeguard. Scores are calculated regardless of whether invariant unit structure exists. Rasch measurement, by contrast, conditions the existence of measurement on empirical conformity to the model. Without model fit, there is no measure.

Interpreting Rasch possession therefore requires understanding three linked properties: logits represent invariant quantitative differences; statistical inference can be applied to evaluate differences in possession; and model fit ensures that measurement validity is empirically tested. Together, these features establish Rasch measurement as a lawful quantitative framework for evaluating therapy impact in latent constructs.

INTERPRETING RASCH POSSESSION CHANGE: CHANGE IN QUANTITY, NOT CHANGE IN SCORE

When Rasch measurement is used to evaluate therapy impact, the outcome of interest is not a change in score but a change in possession of a latent attribute. This distinction is decisive. In psychometric scoring systems, pre–post comparisons yield differences in summed ordinal responses. These differences are expressed numerically, and statistical tests may indicate significance. However, because the underlying scores lack invariant unit structure, a statistically significant difference represents a change in scoring, not necessarily a change in quantity. The numerical difference has no guaranteed quantitative meaning. It may reflect measurement noise, scaling artifacts, or arbitrary scoring conventions.

In Rasch measurement, the situation is fundamentally different. The Rasch model transforms ordinal responses into measures expressed on a logit ratio scale. Each person's location on the latent continuum is estimated with an associated standard error. A pre–post intervention comparison therefore yields two logit measures—each representing quantified possession of the

latent trait—and two standard errors reflecting measurement precision. The difference between these logit values represents a difference in quantity, not merely a difference in score.

Statistical significance in this framework tests whether the observed change exceeds what would be expected from measurement error. The change in logits is divided by the combined standard error to determine whether the shift is statistically reliable. If the resulting statistic exceeds the critical threshold, the improvement is considered statistically significant. Crucially, this means that the individual or group has demonstrated a real change in quantified possession of the latent attribute beyond random variation.

Because the logit scale possesses invariant unit structure, the magnitude of change is interpretable. A 0.5 logit improvement reflects a defined shift in the probability structure linking persons and items. A 1.0 logit shift corresponds to a substantial change in the odds of endorsing more severe or difficult items. These changes are not artifacts of scoring; they represent movement along a quantitative continuum defined by the measurement model. Thus, statistical significance in Rasch measurement indicates a genuine change in quantity.

This stands in sharp contrast to psychometric scoring. In summed-score systems, differences between pre and post totals may reach statistical significance, especially in large samples, but the underlying unit is undefined. A five-point change does not represent five units of a measurable attribute; it represents five increments in an ordinal scoring convention. Without invariant units, arithmetic operations lack empirical interpretation. Statistical significance may detect change in scores, but it cannot confirm change in quantity.

Interpreting possession change in Rasch measurement therefore restores the logical order of scientific evaluation. Measurement is first established through model fit and invariance. Change is then evaluated quantitatively in logit units. Statistical testing assesses whether that change is real. The result is a coherent framework in which therapy impact is expressed as change in measurable quantity, not merely change in numerical score.

RASCH IS REPRESENTATIONAL MEASUREMENT

This requirement aligns Rasch measurement with representational measurement theory. Measurement exists only when numerical assignments preserve empirical relations. The Rasch model establishes this correspondence. The probability structure defined by the model ensures that numerical assignments reflect empirical relations between persons and items. Measurement emerges from empirical structure rather than scoring conventions.

The Rasch model also resolves the problem of invariance across populations. Because the measurement unit is defined by the model, measures remain invariant across samples. This property allows meaningful comparison between individuals, populations, and time points. Psychometric scores do not provide this invariance. Their meaning depends on the specific instrument and population.

Another essential feature of Rasch measurement is its falsifiability. Measurement claims can be tested. If observations do not conform to the model, measurement has not been achieved. This

property aligns Rasch measurement with the principles of normal science. Measurement models must be empirically validated. Psychometric scoring lacks this requirement. Scores are accepted without demonstrating measurement properties.

The Rasch model also distinguishes between measurement and statistical description. Statistical analysis describes relationships between variables. Measurement establishes quantities. The Rasch model provides measurement. Statistical analysis can then be applied to measures. Without measurement, statistical analysis operates on numerical descriptions rather than quantities.

The implications for therapy impact evaluation are profound. Latent attributes such as pain severity, functional impairment, and quality of life cannot be measured through summated scores. These scores do not possess invariant unit structure. Arithmetic operations performed on scores do not correspond to empirical operations. Therapy impact claims based on scores lack measurement validity.

Rasch measurement provides the only mechanism for transforming ordinal responses into quantitative measures. These measures possess invariant unit structure. Differences between measures represent differences in the latent attribute. Therapy impact can therefore be quantified. Changes in logit measures correspond to empirical changes in the possession of the latent trait. Therapy impact claims become falsifiable and replicable.

The Rasch model also allows meaningful comparison between therapies. Because measures possess invariant unit structure, differences between therapies represent differences in quantity. This property is essential for evaluating therapy effectiveness. Psychometric scores do not provide this capability. Differences between scores may reflect scoring conventions rather than empirical differences.

The Rasch model also aligns measurement with arithmetic. Arithmetic operations require quantities. Rasch measurement produces quantities. Arithmetic operations performed on logit measures possess empirical meaning. This alignment restores the logical order of measurement preceding arithmetic.

The Rasch model therefore resolves the fundamental failure of psychometrics. It transforms ordinal observations into quantitative measures. It establishes invariant unit structure. It satisfies the axioms of representational measurement theory. It aligns measurement with arithmetic. It provides the only scientific framework for measuring latent attributes. The distinction between psychometric scoring and Rasch measurement is absolute. Psychometric scoring produces numerical descriptions. Rasch measurement produces quantities. Therapy impact evaluation requires quantities. Psychometric scores cannot satisfy this requirement. Rasch measurement provides the only lawful basis for latent trait measurement.

This conclusion is not methodological but logical. Measurement requires invariant unit structure. Rasch measurement provides this structure. Psychometric scoring does not. The existence of latent trait measurement depends on Rasch transformation.

SECTION IV

RECONSTRUCTING HEALTH TECHNOLOGY ASSESSMENT: FROM PSYCHOMETRIC SCORING TO MEASUREMENT-BASED SCIENCE

The recognition that psychometric scoring does not constitute measurement leads to an unavoidable conclusion: health technology assessment, in its present form, does not measure therapy impact when that impact is expressed in latent constructs such as pain severity, functional impairment, or need fulfillment. Instead, it operates within a scoring paradigm in which numerical assignments are treated as quantities without satisfying the axioms required for measurement. This inversion of scientific logic has profound consequences. It renders therapy impact claims structurally non-quantitative while presenting them in numerical form. The reconstruction of HTA therefore requires a return to first principles, beginning with the requirement that measurement must precede arithmetic.

FORMS OF MEASUREMENT

The essential reconstruction begins by recognizing that there are only two lawful forms of measurement applicable to therapy evaluation. The first is the linear ratio scale applicable to manifest attributes. Manifest attributes are directly observable and possess a natural zero. Time, survival duration, hospital days, and resource utilization are examples. These attributes possess invariant unit structure. Arithmetic operations performed on such measures correspond directly to empirical operations. Therapy impact expressed as additional survival time, reduction in hospital days, or reduction in resource utilization satisfies the requirements of measurement. These claims are falsifiable, replicable, and capable of supporting cumulative scientific knowledge.

The second lawful form is the logit ratio scale produced through Rasch measurement. Latent attributes such as pain severity, symptom burden, or functional limitation cannot be observed directly. They must be measured through transformation of ordinal observations into invariant instrument-based quantitative measures. Rasch measurement achieves this transformation. The resulting logit scale possesses invariant unit structure and a defined zero point. Differences between logit values represent empirical differences in the latent attribute. Therapy impact claims expressed as changes in logit measures therefore possess quantitative meaning.

NUMERICAL STORYTELLING

This distinction between manifest and latent measurement structures resolves a problem that has plagued HTA since its inception. Psychometric scoring attempts to bridge the gap between ordinal observations and quantitative claims through summation and statistical manipulation. These operations cannot create measurement. They produce numerical descriptions without invariant unit structure. The resulting claims cannot support arithmetic operations. When such scores are combined with manifest measures such as time, the resulting composite constructs possess no lawful mathematical properties.

The QALY illustrates this failure. It multiplies time, which is a ratio measure, by utility scores, which are ordinal or at best interval approximations derived from psychometric scoring. Because

utility scores lack ratio scale properties, the resulting product cannot be interpreted as a quantity. The QALY therefore does not measure anything. It produces numerical artifacts that resemble quantities but lack the mathematical properties required for measurement. This failure does not reflect a technical limitation but a logical impossibility. Arithmetic operations require quantities. Scores are not quantities.

Reconstruction of HTA requires abandoning composite constructs that lack dimensional homogeneity. Therapy impact claims must instead be expressed in terms of lawful measurement structures. Manifest attributes must be measured on linear ratio scales. Latent attributes must be measured on Rasch logit ratio scales. This requirement restores the logical foundation of measurement. Arithmetic operations regain empirical meaning. Therapy impact claims become scientifically evaluable.

This reconstruction also transforms the role of protocols in HTA. Instead of generating modeled simulations based on composite constructs, protocols must specify measurable endpoints. For manifest attributes, this involves defining observable quantities such as survival duration, time to clinical events, or resource utilization. For latent attributes, protocols must employ Rasch-validated instruments capable of producing invariant logit measures. Therapy impact claims then become statements about measurable changes in quantity.

FALSIFIABILITY

This transformation restores falsifiability. Measurement-based claims can be empirically tested. Observations either provisionally confirm or refute the claim. Composite constructs derived from scoring cannot be falsified because they do not represent quantities. They function as numerical narratives rather than empirical claims. Reconstruction of HTA therefore restores its status as a scientific discipline.

The implications extend beyond methodology. Measurement establishes the possibility of cumulative knowledge. Quantities can be compared across studies, populations, and time. Therapy impact claims expressed in measurement units can be replicated and refined. Psychometric scores do not support cumulative knowledge because their meaning depends on scoring conventions. Measurement provides continuity. Scoring provides only numerical description.

This reconstruction also simplifies HTA. The complexity of simulation modeling arises from the attempt to compensate for the absence of measurement. When therapy impact is expressed in lawful measurement units, simulation becomes unnecessary. Therapy impact can be observed directly. Claims can be evaluated empirically. The elaborate modeling structures that currently dominate HTA reflect the absence of measurement rather than its presence.

The transition to measurement-based HTA does not require abandoning the evaluation of latent constructs. It requires measuring them correctly. Rasch measurement provides the necessary framework. Instruments must be constructed and validated according to Rasch principles. The resulting logit measures provide quantitative representation of latent attributes. Therapy impact claims expressed in logit units possess empirical meaning.

RECONSTRUCTION

This transition also aligns HTA with the standards of other scientific disciplines. Physics, chemistry, and engineering rely on invariant measurement structures. Quantitative claims in these disciplines are meaningful because they are grounded in measurement. HTA can achieve the same status only by adopting lawful measurement structures. Psychometric scoring cannot satisfy this requirement.

The consequences of failing to reconstruct HTA are equally clear. Continued reliance on psychometric scoring will perpetuate the production of numerical artifacts that lack empirical meaning. Therapy impact claims will remain structurally detached from measurement. Scientific credibility will erode as the absence of measurement becomes increasingly apparent. The field will remain within a closed numerical framework incapable of producing cumulative knowledge. It will continue to live within the false global HTA measurement memeplex.

Conversely, reconstruction based on representational measurement theory restores scientific legitimacy. Therapy impact claims become measurable quantities. Empirical evaluation becomes possible. Knowledge becomes cumulative. HTA transitions from numerical storytelling to measurement-based science. HTA finally adopts the standards of the scientific revolution.

This transition is not optional. It reflects the logical requirements of measurement itself. Arithmetic cannot precede measurement. Quantities cannot be created through scoring conventions. Measurement must be established before quantitative claims can be made. Rasch measurement provides the only framework capable of measuring latent attributes. Linear ratio measurement provides the framework for manifest attributes. Together, these two measurement structures provide the complete foundation for therapy impact evaluation.

The reconstruction of HTA therefore represents a return to scientific first principles. It replaces scoring with measurement. It replaces numerical description with quantitative representation. It restores the logical order of measurement preceding arithmetic. It aligns HTA with the standards that govern all quantitative science. Psychometrics promised measurement but delivered scoring. Representational axioms of measurement deliver measurement. The future of HTA depends on recognizing this distinction and reconstructing its evaluative framework accordingly.

CONCLUSION

The argument developed across these four sections leads to a conclusion that is both unavoidable and transformative. Psychometrics is not measurement. It is a system of numerical scoring that assigns numbers to observations without establishing the invariant quantitative structure required for arithmetic operations. For decades, health technology assessment has relied on psychometric scoring systems as if they produced measurable quantities. Composite indices, summed ordinal responses, and preference-weighted utilities have been treated as quantitative measures of therapy impact. Yet these constructs lack dimensional homogeneity, lack invariant unit structure, and lack the true zero required for ratio measurement. They are numerical descriptions, not measurements.

This distinction is not semantic. It is structural. Measurement is defined by the existence of invariant quantitative relations between empirical attributes and numerical representation. Without these relations, arithmetic operations are invalid. Multiplication, division, and aggregation require quantities. Scores do not satisfy this requirement. The consequence is that many of the central constructs in HTA, including utilities and QALYs, cannot support the arithmetic operations upon which cost-effectiveness claims depend. These claims therefore lack the mathematical and empirical foundation required for scientific evaluation.

The Rasch model resolves this problem by providing the only lawful framework for transforming ordinal observations of latent attributes into quantitative measures. Through conjoint simultaneous measurement, Rasch analysis produces invariant logit ratio scales that satisfy the axioms of representational measurement. These logit measures represent quantities. Therapy impact claims expressed in logit units possess empirical meaning, are falsifiable, and support cumulative knowledge. In parallel, manifest attributes such as survival time and resource utilization must be measured on linear ratio scales. These two measurement structures—linear ratio scales for manifest attributes and Rasch logit ratio scales for latent attributes—exhaust the lawful forms of measurement applicable to therapy evaluation.

The reconstruction of HTA therefore requires abandoning psychometric scoring as a basis for quantitative claims and adopting lawful measurement structures as its foundation. This transition restores the logical order of scientific inquiry: measurement preceding arithmetic, quantity preceding calculation, and empirical observation preceding numerical representation. It transforms HTA from a discipline dependent on numerical convention into one grounded in empirical measurement. The choice is clear. Either HTA continues to operate within a scoring paradigm that produces numerical artifacts, or it embraces measurement and secures its future as a scientific discipline capable of producing evaluable and replicable knowledge.

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.

A NEW START IN MEASUREMENT FOR HEALTH TECHNOLOGY ASSESSMENT

For readers who are looking for an introduction to measurement that meets the required standards, Maimon Research has just released two distance education programs. These are:

- Program 1: Numerical Storytelling – Systematic Measurement Failure in HTA.
- Program 2: A New Start in Measurement for HTA, with recommendations for protocol-supported claims for specific objective measures as well as latent constructs and manifested traits.

Each program consists of five modules (approx. 5,500 words each), with extensive questions and answers. Each program is priced at US$65.00. Invitations to participate in these programs will be distributed in the first instance to 8,700 HTA professionals in 40 countries.

More detail on program content and access, including registration and on-line payment, is provided with this link: https://maimonresearch.com/distance-education-programs/

REFERENCES

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

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

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

[4] Bond T, Z Yan, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 4th ed. New York: Routledge, 2021

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

[6] Pigliucci M. Nonsense on Stilts: How to Tell Science from Bunk. Chicago: University of Chicago press, 2010

[7] Popper K. Objective Knowledge: An Evolutionary Approach. Revised edition. Oxford: Oxford University Press, 1979

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