MAIMON WORKING PAPER No 15

VALUE CLAIM PROTOCOL: RASCH, LATENT CONSTRUCTS AND ATTRIBUTE POSSESSION

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

ABSTRACT

Health Technology Assessment (HTA) must rigidly adhere to the axioms of representational measurement. These mandate the exclusive use of linear ratio or Rasch logit ratio measures—unidimensional, invariant, and quantitatively meaningful. Multiattribute and generic preference-based tools like EQ-5D-3L, and the QALY framework built on non-linear composites and ordinal scores, fail these criteria and thus must be abandoned. Instead, HTA must undergo a transformative renewal: replacing legacy instruments with rigorously constructed Rasch-compliant tools.

A latent construct is not measured per se; rather, a specific, clearly defined attribute, such as needs-fulfillment, is operationalized and measured. The Rasch measurement model provides the only mathematically sound method to convert bounded, ordinal responses into a linear, invariant logit scale. This scale, with its constant relative differences, enables robust arithmetic operations and supports valid inference about therapy impact. Though it lacks a classical "true zero," the logit's designated zero as the 50% probability point yields a ratio-level metric aligned with measurement principles.

Instrument development under the Rasch framework involves qualitative item construction, empirical targeting, and iterative calibration of difficulties and respondent abilities on the same logit continuum. This process ensures unidimensionality, local independence, and precision, especially around the 0-logit region, where most therapeutic changes occur.

Logit-based scores allow conventional statistical analysis, group-level comparisons, and meaningful interpretation of change over time. Rescaling (e.g., via USCALE and UMEAN) enhances interpretability without compromising measurement integrity. Crucially, despite the availability of accessible Rasch software tools, the HTA field has largely failed to utilize this methodology beyond calibration. Instruments are often abandoned at development, and users revert to summing ordinal scores, undermining scientific validity.

Manufacturers committed to scientific integrity must thus embrace Rasch measurement from construct conceptualization through instrument development and change interpretation. This commitment positions them as leaders in renewing outcomes measurement, even as it challenges the current HTA and regulatory paradigms. Only by anchoring value claims in scientifically defensible, Rasch-aligned measures can HTA regain credibility. Rasch measurement is not optional, —it is the essential foundation for transparent, evaluable, and replicable claims about therapy response.

INTRODUCTION

It has been emphasized in a number of Maimon Working Papers that the axioms of representational measurement impose strict requirements on the evaluation of therapeutic impact in health technology assessment (HTA) [1] Only two types of measures are acceptable: linear ratio measures with constant absolute differences, and Rasch logit ratio measures with constant relative differences [2]. These standards are not peculiar to HTA, they are foundational across the physical sciences and the more mature branches of the social sciences. For any latent construct relevant to disease management or target patient populations, such as need fulfillment, the required standard is clear: the measure must be unidimensional, linear, a Rasch logit ratio scale, and invariant. There are no exceptions.

The implications for HTA are far-reaching. Multiattribute, generic preference-based instruments such as the EQ-5D-3L fail to meet these standards and must be rejected. So too must the entire edifice of QALY-based modeling and the accompanying reference case simulations, which are built on assumptions that ignore these fundamental principles. Likewise, all disease-specific instruments currently in use are fatally compromised if they do not demonstrate Rasch compliance. The failure to adhere to the axioms of measurement theory has resulted in four decades of misguided practices, underpinned by an uncritical acceptance of ordinal scores, non-linear composites, and mathematically indefensible models.

This situation is analogous to the mythic task of cleansing the Augean stables, a monumental cleanup of accumulated error and misinformation. Yet despite the clarity of measurement theory, HTA continues to exist in a parallel relativistic universe, dominated by numerical storytelling. The leadership of the HTA community, academic, professional, and regulatory, has largely chosen to ignore the standards of normal science in favor of constructing imaginary value claims through simulation and assumption-driven models. This is not science; it is a performance of pseudoscientific reasoning justified by convenience, cost, and inertia.

If this state of affairs appears absurd, then the only remedy is a program of fundamental renewal. HTA must realign itself with the standards that define credible, evaluable, and replicable science. This requires abandoning defective legacy instruments and rebuilding our assessment tools on the firm foundation of Rasch measurement. Only then can HTA reclaim scientific legitimacy.

This reformation has immediate implications for a value claim protocol framework. If manufacturers intend to submit, or formulary committees demand, claims based on measured subjective outcomes, then those claims must be grounded in Rasch-structured instruments. There is no middle ground. Adherence to Rasch measurement is not a methodological preference, it is a precondition for scientific integrity in evaluating therapy impact for latent construct manifestations.

THE LATENT CONSTRUCT

It is important to emphasize from the outset that it is not the latent construct per se that is the object of measurement, but rather a specific property or attribute of that construct that is of interest. A latent construct can be understood, ontologically, as a hypothesized, unobservable entity that exists within a conceptual domain and cannot be directly apprehended or measured. Constructs such as "patient engagement," "treatment satisfaction," or the broader "patient voice" represent real phenomena in the lived experience of patients, but their existence is inferred rather than directly observed. These constructs are postulated to explain patterns in human behavior or experience and are accessed only through the development of observable indicators.

However, latent constructs may admit of multiple associated properties, and it is a single, well-defined property or attribute, such as needs fulfillment as an attribute of the patient voice, that becomes the actual object of measurement. We do not measure the patient voice in general; rather, we aim to assess whether and to what extent a patient exhibits or experiences a particular attribute of that construct. The evolution of objective knowledge in this context refers to the increasing refinement and empirical understanding of the structure, manifestation, and measurement of such properties. That is, progress is made not by arbitrarily expanding descriptive frameworks, but by isolating specific attributes of latent constructs and subjecting them to rigorous standards of fundamental measurement.

The Rasch measurement model provides the only mathematically rigorous framework for this task. It focuses on the development of instruments, typically questionnaires, designed to yield unidimensional, linear, Rasch logit ratio, and invariant measures of the attribute in question. Rasch modeling is not about scoring or summing ordinal responses; rather, it is about constructing a measurement structure that aligns person ability with item difficulty on a single scale, thereby making possible a valid inference about possession of the attribute. The model operationalizes the process of *manifestation*, in which an observable response pattern is taken to reflect the level of an underlying property of a latent construct.

THE INEVITABLE LOGIT

The Rasch model is not simply a convenient method among others; it is the only approach consistent with the axioms of representational measurement that allows the transformation of bounded, ordinal, subjective responses into a meaningful, quantitative scale. In the context of health technology assessment, where value claims may rely on subjective patient experiences, this transformation is indispensable.

Subjective responses, such as those relating to a dimension of quality of life, symptom relief, or perceived well-being, are typically reported using bounded rating scales, often with arbitrary endpoints. These raw scores are ordinal, meaning they can only express relative rankings but provide no information about the distance between points. Such scales cannot be assumed to have either a true zero or equal intervals, and they are invariably constrained by their format, making them unsuitable for arithmetic operations such as multiplication or division. A scale with constant absolute differences, a linear interval or ratio scale, requires a zero point that is not assigned but true, indicating the complete absence of the property being measured. This is impossible to

establish for subjective experiences, which are inherently continuous, context-sensitive, and resistant to absolute delimitation.

The Rasch logit scale circumvents this problem by transforming ordinal data into an interval scale based on the probabilistic relationship between person ability (or trait level) and item difficulty. This transformation yields a unidimensional, linear, and invariant measure on a log-odds (logit) scale. Although the scale lacks a true zero in the classical sense, it does possess an *assigned zero,* a logit value of zero, defined as the point at which the likelihood of success (or affirmation of an item) is 50%. Importantly, the logit scale has *constant relative differences*, meaning that a one-logit difference represents the same proportional change in odds, regardless of position on the scale. This characteristic makes the scale ratio in form and supports the full range of arithmetic operations necessary for measurement, even in the absence of constant absolute differences.

In this sense, the Rasch logit ratio scale is not a compromise, but the only legitimate path to constructing scientifically valid measures from subjective data. It provides a means to anchor patient-reported outcomes in the framework of fundamental measurement by ensuring that claims about therapy impact rest on a stable, interpretable, and replicable metric. Any attempt to construct or interpret such claims without this transformation risks producing artifacts rather than meaningful data, undermining the scientific integrity of HTA. Only Rasch measurement offers the structure required to turn subjective experiences into quantitative evidence that supports evaluable and replicable value claims.

RASCH INSTRUMENT DEVELOPMENT

Instrument development for subjective responses is made relatively straightforward through the application of Rasch measurement principles. With access to any one of several Rasch modeling software packages, the process of calibrating responses to generate a unidimensional, linear, Rasch logit ratio scale is computationally efficient, often taking only a few minutes once the item pool and dataset are in place. However, while the mechanical process may appear straightforward, the more challenging and critical stage lies in the identification, development, and refinement of items that are proposed to populate the measurement instrument.

Items are not chosen arbitrarily nor adapted from existing patient-reported outcome (PRO) instruments, which frequently fail to meet the standards of fundamental measurement. Instead, items must be drawn from rigorous qualitative inquiry, typically through structured and in-depth interviews with representative members of the target patient population. The purpose of these interviews is to generate items that reflect different levels of difficulty, or, more precisely, different thresholds for endorsing the attribute of interest. Each item must capture a distinct point along the underlying latent trait continuum, representing degrees of possession of the attribute such as needs fulfillment or symptom burden.

What sets Rasch instrument development apart from traditional PRO approaches is its unique commitment to the conjoint simultaneous measurement of persons and items. Within a Rasch framework, both the ability of the respondent (their position on the latent trait) and the difficulty of the item (the likelihood that the item will be affirmed) are mapped onto the same logit interval scale. The result is a scale that supports not only ranking but also measurement, with the units

representing constant relative differences along the latent trait. Respondent ability and item difficulty are expressed in the same metric, making possible valid comparisons, predictions, and interpretation of change.

A core feature of the Rasch model is its iterative structure. Initial item selection is followed by data collection and model fitting, during which the distribution of respondent abilities is evaluated against the distribution of item difficulties. Misfitting items, those that show inconsistent responses across the range of respondent abilities, are flagged by the software and subject to review. This review may result in item modification, removal, or replacement. The objective is to achieve a well-targeted scale where items span the full range of respondent abilities and the model fit statistics (infit and outfit mean squares) confirm that the assumptions of unidimensionality and local independence are satisfied.

This iterative process is central to the scientific legitimacy of Rasch measurement. It ensures that the final instrument does not merely reflect statistical convenience but aligns with the underlying construct and the measurement properties required by the axioms of fundamental measurement. Only by this route can subjective responses be transformed into ratio-level evidence to support evaluable and replicable value claims in health technology assessment.

DISTRIBUTION OF ITEM DIFFICULTIES

When constructing a Rasch instrument, particular attention must be paid to the distribution of item difficulties along the latent trait continuum. The Rasch model assumes that meaningful measurement occurs when there is good alignment, or "targeting" between the distribution of item difficulties and the distribution of respondent abilities. If items are not appropriately targeted, measurement precision can be compromised, particularly in the central range of the scale where small changes in logit scores often correspond to clinically meaningful differences in patient outcomes.

A common feature of Rasch measurement is that the mean item difficulty is set at 0 logits by convention. This allows respondent logit scores to be interpreted relative to the average item difficulty. However, this does not imply that all items must cluster exactly at 0; rather, what is required is a distribution of item difficulties that adequately spans and centers around the expected range of respondent abilities. For many clinical applications, especially those involving interventions that yield modest improvements in attribute possession, it is crucial that sufficient item density exists in the logit range from approximately –1.0 to +1.0.

The reason for this is mathematical and interpretive. Logit values express the log-odds of endorsing more difficult items. Because the logit scale is nonlinear when transformed into probabilities, the percentage change associated with a shift from –1.0 to +1.0 logits is much smaller in absolute terms than changes at more extreme ends of the scale. For example, moving from –1.0 logits to 0 logits increases the probability of affirming a given item from about 27% to 50%, while moving from 0 to +1.0 logits increases it further to roughly 73%. Although this represents a doubling in odds, the associated percentage increase (from 27% to 73%) is bounded and nonlinear, emphasizing how subtle shifts in the logit range near 0 can still represent significant gains in perceived function or well-being.

To detect and interpret these small but meaningful shifts reliably, the Rasch instrument must contain items with difficulties distributed around the 0-logit mark. Without sufficient item coverage in this central region, changes in patient scores may be poorly estimated or inadequately supported by the data, reducing the sensitivity of the instrument to detect therapy effects. This is especially important in longitudinal or intervention studies where we aim to track incremental improvements.

Additionally, items far removed from the respondent distribution (i.e., much easier or harder than most respondents' ability levels) contribute little to the precision of measurement in the target population. Including too many such items leads to local gaps and imprecision in the range where change is most likely to occur. For this reason, item selection must be empirically guided to ensure that the measurement scale is populated densely where most patients are expected to fall, and where therapy is most likely to produce an effect.

Thus, clustering items around 0 logits is not arbitrary but fundamental. It ensures the Rasch scale is sensitive in the zone where most patient-level changes are observed, preserving the interpretability and evaluability of value claims based on subjective therapy response.

INTERPRETING THE RASCH LOGIT MEASURE

A Rasch instrument or questionnaire typically consists of 25 to 30 items, depending on the complexity of the latent construct and the breadth of the attribute being measured. These items can employ either dichotomous response formats (e.g., yes/no) or polytomous formats (e.g., Likert-type scales with ordered categories). Regardless of format, what distinguishes Rasch-based measurement is not the raw score or count of item responses, as is common in traditional scoring models, but the pattern of those responses and their alignment with the probabilistic expectations of the Rasch model.

It is critical to emphasize that response to therapy is not defined by an accumulation of responses across items, as in traditional scoring models, but by the shifting distribution of individual responses over time. Each respondent's pattern of responses, across the full set of items, is analyzed to estimate their position on the latent trait continuum. This is achieved using an iterative maximum likelihood estimation procedure, which identifies the logit score that best represents the respondent's possession of the attribute, given their response pattern and the calibrated difficulty of each item.

Each individual logit score represents a location on the Rasch scale, measured in log-odds units, that expresses the relative probability of affirming more difficult items. Once logit possession scores are calculated for each respondent, they can be averaged to obtain a group-level estimate of possession of the latent trait, whether at baseline, after intervention, or across multiple time points. This average logit score can then be compared using conventional statistical methods to evaluate therapy response.

Because the Rasch logit scale is a ratio scale with constant relative differences, standard arithmetic operations such as calculating means, differences, standard deviations, and conducting statistical tests are appropriate. For example, suppose we observe a mean group logit score of –0.5 before

therapy and +1.5 after therapy. With a sample size of 200 respondents and a common standard deviation of 0.2 logits, the effect size (Cohen's d) is calculated as:

Effect size = (1.5 – (–0.5)) / 0.2 = 2.0 / 0.2 = 10.0

This extremely large effect size indicates a substantial shift in possession of the measured attribute. A paired t-test, assuming normality, would yield a test statistic of:

t = (mean difference) / (standard error) = 2.0 / (0.2 / √200) ≈ 2.0 / 0.0141 ≈ 141.4

This test result confirms that the change is statistically significant at any conventional level ($p < 0.001$), assuming model assumptions are satisfied.

Understandably, the audience for claims of possession of a latent trait, either individually or as an average for a target group, will be unfamiliar with logits. While it is possible to transform a logit to a percentage this loses the requirements of the logit scale. The more satisfactory solution is to apply a Rasch transformed score using what WINSTEPS describes as USCALE, which defines how many units on the new scale correspond to one logit and UMEAN which shifts the scale so that the lowest logit aligns with the desired low score; Rasch interval measures are invariant under linear transformation [3] [4].

Consider the following example where the logit range is +/- 2.0 logits. USCALE is estimated from the following:

USCALE = (Desired High Score – Desired Low Score)/Logit range

= (2.0 – (-2.0)) 4 = 0.25

UMEAN = Desired Low Score− (Low Logit × USCALE)

=0 − (−2.0×25.0) = 50.0

Consider an average possession for the group of 0.7 logits. This yields a transformed possession score of 67.5:

Transformed Score = (Logit Value × USCALE) + UMEAN

= (0.7×25.0) + 50.0 = 67.5

The advantages of this rescaling are that the transformed scale maintains equal intervals, allowing for accurate measurement of differences in possession. Aligning the scale to a 0–100 range makes it more intuitive for stakeholders, resembling familiar metrics like percentages or standardized test scores. While applying the same USCALE and UMEAN values ensures consistency when comparing different test forms or administrations.

While a score of 67.5 on a 0–100 scale derived from Rasch-transformed logits numerically resembles a percentage, it's important to understand that it doesn't represent a percentage in the traditional sense. In classical assessments, percentages often indicate the proportion of correct responses, which are ordinal and don't account for item difficulty or person ability. In contrast, Rasch-transformed scores are interval-level measurements that consider both item difficulty and person ability, providing a more nuanced understanding of performance. Therefore, interpreting a score of 67.5 as a simple percentage of correct answers would be misleading. Instead, it's more appropriate to describe this score as representing a position on a linear scale of possession, where equal intervals reflect equal differences in the underlying construct being measured. This approach maintains the integrity of the measurement and provides more meaningful insights into the assessed trait.

It is of interest to note that with a possession of 0.7 logits the transformed possession score (see above) is 67.5/100 while the corresponding p value logistic transformation is 66.8%. The close numerical values are coincidental due to the specific logit value and rescaling parameters chosen. The logistic transformation is nonlinear and maps logits to probabilities, while the WINSTEPS rescaling is a linear transformation for interpretability. While both methods provide values in a similar range for certain logits, they serve different purposes: The logistic transformation: converts logits to probabilities, useful for interpreting the likelihood of an event compared to a linear rescaling which transforms logits to a user-defined scale for reporting and interpretability, maintaining interval properties.

PROTOCOL CHALLENGES

A manufacturer committed to scientific standards is immediately confronted with a difficult but necessary choice: to recognize that any claim for a therapy's impact based on subjective response must rest on the measurement of a property of a latent construct. There is no scientific alternative. The manufacturer must accept that the only valid path to quantifying therapy benefit in the context of patient-reported outcomes is through Rasch measurement theory. In doing so, they acknowledge that traditional false psychometric approaches, those based on summing ordinal item responses, often from legacy instruments, are incapable of providing interval-level measurement, and therefore incapable of supporting credible or evaluable value claims. Rasch offers not merely an alternative method but the only framework that meets the axioms of fundamental measurement and satisfies the standards of normal science. This decision, however, immediately isolates the manufacturer from standard health technology assessment practices and from the regulatory *status quo*, where almost all subjective claims rest on flawed assumptions and ordinal data misapplied in cost-effectiveness models and QALY frameworks.

The challenge, then, is not one of instrument selection but of conceptual integrity. Rasch measurement requires beginning with a clear definition of the latent construct, a conceptual entity such as treatment satisfaction, needs fulfillment, symptom burden, or functional status. From that definition, the task becomes one of identifying and validating items, questions or statements, that can reliably elicit from respondents a pattern of responses that reflects different levels of possession of a specific attribute of that construct. These items must then be calibrated on a common logit scale with respondent ability, allowing both to be expressed in the same

measurement space. The Rasch model thereby enables the transformation of bounded, ordinal responses into a unidimensional, linear, interval scale; one with constant relative differences and invariant properties, capable of supporting arithmetic operations and change measurement over time.

Yet despite this clarity, very little has been accomplished in practice. Over the past several decades, Rasch measurement has been applied sporadically HTA; usually superficial. In most cases where Rasch methods are invoked, the focus remains entirely on the development of the instrument; the questionnaire itself. Studies report the calibration of items, demonstrate adequate fit statistics, and declare unidimensionality. But the process typically stops there. Researchers seldom proceed to the core purpose of measurement: to estimate and interpret the *level of possession* of the attribute by individuals or groups. Possession scores, expressed in logits, are never reported, and changes in possession following therapeutic intervention are never subjected to proper statistical analysis. Instead, the instrument is handed over to users who then sum raw scores or collapse categories, reverting to ordinal summaries and non-Rasch methods of interpretation. As detailed above, the solution is not to transform logits to percentages but to create a Rasch transformed score. This, on a chosen scale of 0 – 100 is not a percentage but a position on a linear scale of possession.

Several reasons explain this failure to move beyond the instrument itself. The first is conceptual inertia. Most health outcomes researchers have been trained within the framework of classical test theory, where the emphasis is on reliability coefficients and aggregate scores. The language and assumptions of Rasch measurement are unfamiliar, and there is widespread misunderstanding of its implications. A second factor is institutional resistance. Regulatory bodies and HTA agencies have not required Rasch-based evidence, and therefore manufacturers are not incentivized to go beyond what is currently accepted. As long as flawed instruments like the EQ-5D or SF-36 are permitted in submissions and economic models, the cost and effort of developing Rasch instruments that support ratio-level claims appear unjustified in the short term. There is also a more practical issue: the absence of expertise. While Rasch software is readily available and instrument calibration is technically accessible, the correct interpretation of logit scores, proper targeting of items to expected trait distributions, and valid expression of changes in possession require a level of training and commitment that is rare in the applied outcomes research community.

Even where the few Rasch instruments are available, their structure is frequently undermined by poor implementation. Developers may calibrate a scale but fail to ensure that items are appropriately distributed around the expected range of respondent abilities. If items cluster too high or too low, respondents will generate extreme scores that cannot be meaningfully interpreted or tracked over time. More importantly, very few studies carry through to the most critical phase: evaluating whether an observed change in logit scores represents a statistically significant and clinically meaningful shift in possession. This is where Rasch's strength lies: in establishing that an intervention has altered a measurable attribute of a latent construct in a replicable and interpretable way. Yet the field remains stuck in the initial stages, treating Rasch as a statistical curiosity rather than the only standard for scientific latent construct claim validation.

For a manufacturer who sees Rasch as the only viable pathway to credible latent trait value claims, this landscape is discouraging but not insurmountable. The opportunity exists to lead a transformation in outcomes measurement by refusing to endorse surrogate metrics or simulation-

based assumptions. By committing to the full Rasch pathway, from construct definition to possession estimation and change, a manufacturer can generate evidence that is not only scientifically credible but replicable and testable in future studies. While this may currently place them outside the dominant HTA paradigm, it positions them at the leading edge of a renewal in health outcomes research that reclaims the principles of measurement as the foundation of evidence. Until such leadership is taken seriously, Rasch will remain underused, and health technology assessment will continue to rest on claims that cannot, in principle, be measured.

Despite real or imagined challenges, properly selecting items that fit the Rasch model ensures it satisfies the foundational axioms of representational measurement in probabilistic terms. This alignment endows the model with specific objectivity, meaning that comparisons between individuals remain independent of the particular items used, and vice versa, provided the data adhere to the model While Krantz et al.'s framework is grounded in deterministic axioms of conjoint measurement, Rasch extends this by operationalizing latent traits probabilistically, enabling empirical testing and fit evaluation In doing so, the Rasch model achieves a uniquely rigorous form of measurement: it preserves invariant measurement through sufficiency of statistics and supports robust estimation using conditional maximum likelihood, in contrast with purely algebraic models By shaping measurement theory in this way, the Rasch model stands as a powerful, inherently probabilistic and empirically grounded solution to the challenge of latent-trait quantification.

CONCLUSIONS

While experience with the Rasch logit ratio measure as the only legitimate measure of latent construct properties or attributes is limited in HTA, there has to be a breakthrough to make clear the unique role of Rasch measurement. Persisting with legacy measures either represents an admission of defeat or a degree of satisfaction with a status quo that rests on the patent lack of understanding of fundamental measurement by the target audience. The protocol must fill two functions. First, an educational function in explaining why Rasch is the only option for a measure of therapy response and, second, how a proposed instrument has been developed to meet an unmet need for accurate therapy response claims.

This dual purpose is essential. The audience for HTA submissions is rarely familiar with the principles of fundamental measurement, and even less so with the implications of Rasch modeling. Many still believe that summing ordinal responses and calculating mean scores from bounded, non-linear scales provides meaningful results. The protocol must clearly state that only Rasch transforms such data into interval-level evidence capable of supporting arithmetic operations and change measurement. It must also demonstrate that the instrument has not merely been calibrated but designed to capture real changes in the possession of the attribute across a defined patient population. Without this step, there is no basis for claiming impact. The Rasch protocol is not just a method; it is a scientific position grounded in a measurement framework that allows for reproducible, evaluable, and replicable claims. Without it, HTA continues to traffic in proxies and preferences, not evidence.

The limited attention given to the imperative of Rasch measurement in health technology assessment is puzzling, particularly given the accessibility of well-established software packages and the routine use of Rasch methods in other disciplines such as education, psychology, and

rehabilitation science. Tools such as Winsteps, RUMM2030, and ConQuest have made it technically straightforward to construct instruments that meet the axioms of fundamental measurement, transforming ordinal data into unidimensional, interval-level scales. In contrast, the health outcomes field has largely remained tied to outdated disease specific instruments, relying on ordinal summed score data that cannot support valid arithmetic operations or meaningful claims of change. This resistance may be rooted in both institutional inertia and a widespread lack of awareness of the principles of Rasch modeling. Furthermore, regulatory and HTA environments have failed to demand rigorous measurement standards, allowing the continued use of composite scores and preference-based indices that violate the requirements of measurement theory. The result is a culture of numerical storytelling that privileges convenience over scientific credibility. Until this belief system shifts, and Rasch measurement is recognized not as an optional refinement but as a necessary foundation for credible value claims, HTA will continue to produce outcomes that are neither replicable nor interpretable in a scientifically meaningful way.

ACKNOWLEDGEMENT

The development of this paper has benefited from the use of AI-assisted tools, specifically ChatGPT by OpenAI (https://chat.openai.com/), Version 5, for tasks including revised drafting and text editing. The author takes full responsibility for the content and any errors that may remain

REFERENCES

[1] Krantz D, Luce R, Suppes P, Tversky A. *Foundations of Measurement, Volume I: Additive and Polynomial Representations*. San Diego and London: Academic Press, 1971.

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

[3] Wright B, Linacre J. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil*. 1989;70(12):857–860

[4] Linacre J. Investigating rating scale category utility. *J Outcome Measurement*. 1999;3(2):103–122.