MAIMON WORKING PAPER No. 20 SEPTEMBER 2025

THE FAILURE OF ICHOM: PART 2 A STRATEGY FOR SURVIVAL

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

ABSTRACT

The International Consortium for Health Outcomes Measurement (ICHOM) has positioned itself as a leader in standardizing outcomes across disease areas by promoting "standard sets" of recommended instruments. The ambition, to create a global language for value-based care, is laudable, but the method is fatally flawed. ICHOM has elevated legacy instruments that fail the most basic requirements of representational measurement theory. These instruments, often patient-reported outcome questionnaires, collapse multiple attributes into ordinal totals that are then presented as if they were interval or ratio measures. They lack unidimensionality, additivity, and invariance; they cannot sustain meaningful arithmetic or falsifiable claims. The result is not science but numerology with the appearance of rigor.

This paper asks a sharper question than the first critique: can ICHOM survive if it continues on this trajectory, or must it pivot to science? The answer is clear. There are only two admissible forms of measurement for credible health outcomes: linear ratio scales for manifest attributes and Rasch logit ratio scales for latent constructs. All other devices including summed scores, preference utilities, composite indices fail the axioms of measurement and collapse on inspection. Unless ICHOM adopts these standards, it cannot deliver claims that are credible, evaluable, or replicable.

The pivot required is radical but achievable. ICHOM must transform itself from a catalogue of consensus instruments into a curator of admissible measures. Every endorsed claim must be tied to a protocol specifying the attribute, the scale type, the timeframe, and the test of invariance. Manifest attributes such as hospital days or meters walked must be measured on ratio scales with true zeros and constant units. Latent attributes such as fatigue or dyspnea severity must be modeled on Rasch logit scales, ensuring unidimensionality, ordered thresholds, and invariance across groups. Instruments that cannot meet these standards must be retired or re-engineered.

The stakes could not be higher. For eighty years the requirements of measurement have been known and settled. Stevens, Suppes, Luce, Tukey, the Foundations of Measurement*, Rasch, and Wright made clear the axioms that distinguish numbers from measures. To ignore them is not ignorance but neglect, perpetuating five decades of false claims. ICHOM's survival depends on whether it renounces consensus numerology and embraces measurement. If it does, it can lead a long-overdue reconstruction of health outcomes science. If it does not, it will remain what it is today: a clearinghouse of pseudo-measures, destined to fail.*

INTRODUCTION

The International Consortium for Health Outcomes Measurement (ICHOM) has, since its inception, sought to improve the comparability of outcomes across disease areas by assembling "standard sets" of recommended instruments. The intent is admirable: to bring order to a fragmented field and to create a global lingua franca for value-based care. Yet the method chosen, compiling disease-specific questionnaires on the basis of consensus, rests on a fragile and ultimately indefensible foundation. As argued in Part 1, ICHOM has elevated legacy instruments that fail the basic standards of representational measurement theory (RMT). Most are nothing more than ordinal sums of multiple attributes, presented as if they were interval or ratio measures. They lack unidimensionality, invariance, and additivity. They cannot sustain meaningful arithmetic or generate falsifiable claims. The result is not science but numerology in scientific dress.

This second paper poses a sharper question: can ICHOM survive if it continues on this trajectory, or must it pivot decisively to measurement? The title captures the stakes: *Will ICHOM Survive or Fail?* The answer depends not on its governance structure, its reach, or its brand, but on whether it embraces the only admissible foundations for empirical claims. There are two and only two forms of measurement capable of sustaining evaluable, replicable, and falsifiable claims in health technology assessment: linear ratio scales for manifest attributes and Rasch logit ratio scales for latent constructs [1], All other devices, summed scores, preference utilities, multiattribute composites, fail the axioms of measurement and collapse under scrutiny.

The pivot required is radical but clear. ICHOM must transform itself from a catalogue of disease-specific questionnaires into a curator of admissible measures. Every endorsed outcome claim must be backed by either a linear ratio or a Rasch logit ratio measure. Nothing else will do. This requires a measurement charter that defines admissible endpoints, a filter to exclude non-measures, and a redevelopment program to create Rasch instruments where gaps exist. Only then can ICHOM move from convening consensus to stewarding science.

The argument is not that ICHOM should be perfect or perish. It is that without a pivot to measurement it will certainly perish, reduced to a clearinghouse of nice-looking numbers that cannot bear scientific weight. With a pivot, however, it can survive and indeed lead, becoming the first global body to declare what health outcomes research has so long evaded: that only two types of measures are admissible, and that without them, there is no science and the evolution of objective knowledge[2] .

WHY ICHOM FAILS

ICHOM was launched with the ambition of bringing coherence and comparability to the evaluation of healthcare interventions. Its model has been to convene working groups in specific disease areas, composed of clinicians, academics, and patient representatives, and to assemble "standard sets" of outcomes. These sets are then promoted as the best available metrics for assessing the value of care. At first glance, the approach seems practical, even enlightened. It signals a recognition that outcomes must be defined, collected, and compared, and that consensus is needed if data are to travel across settings. Yet beneath this apparent strength lies the fatal weakness:

ICHOM has built its catalogue on instruments that cannot claim the status of measures. Without measurement, its enterprise collapses.

The root of the failure is the absence of a measurement filter. ICHOM accepts instruments on the basis of consensus, clinical familiarity, or face validity rather than on their ability to meet the RMT [3]. The instruments most often selected are legacy patient-reported outcome measures, questionnaires designed in earlier decades that aggregate responses across domains into summed totals. These totals are treated as if they were interval scores, suitable for arithmetic operations and statistical analysis. In reality they are ordinal counts, reflecting only the rank order of respondents, not the magnitude of differences between them. Summed scores cannot support subtraction or division, cannot sustain claims of improvement or deterioration in quantified terms, and cannot provide the invariance required for fair comparison across individuals or groups [4].

RMT, codified in the 1960s and 1970s, makes explicit what counts as a measure. A measure is a structure-preserving mapping from an empirical relational system to a numerical relational system. This requires unidimensionality, additivity, and invariance. It requires that differences in the numbers correspond to differences in the attribute being measured, and that those differences remain constant across transformations permitted by uniqueness theorems. Summed scores from multiattribute questionnaires fail these requirements in every respect. They combine items from distinct latent constructs, symptoms, functions, moods, into a single number. They presume additivity without demonstrating it. They offer no test of invariance across populations. They produce outputs that look like measures but are not.

The Rasch model, first published in 1960 and elaborated in the following decades, provides the path from ordinal responses to interval measures for latent constructs [5]. Rasch analysis requires that data fit the model rather than the model being tailored to fit the data. It provides a logit scale in which constant relative differences are preserved, items are ordered by difficulty, and persons by ability, all on the same continuum. It enforces unidimensionality and enables invariance testing across groups. If ICHOM were serious about measurement, it would require every latent construct to be assessed with a Rasch-calibrated instrument. It would reject outright any questionnaire that failed to meet these standards. But ICHOM has not adopted this stance. Instead, it has embraced legacy tools whose very structure precludes Rasch conformity.

It is essential to distinguish Rasch measurement from the broader family of item response theory (IRT) models. Rasch is not merely one IRT option among many but a unique framework aligned with the axioms of RMT. Its purpose is to construct a single, unidimensional instrument that locates both persons and items on the same logit scale, thereby quantifying possession of a latent trait. Fit to the Rasch model is a non-negotiable requirement: data must conform to the model, not the other way around, and only then does the resulting scale warrant interval properties, invariance, and additivity. By contrast, IRT, especially in its PROMIS elaborations, is driven by flexibility and model fit in a statistical sense, not by adherence to measurement axioms. PROMIS item banks yield adaptive testing and convenient scoring but do not guarantee unidimensionality or invariant possession profiles. Rasch delivers measurement; IRT delivers prediction without measurement.

This is why ICHOM fails. It does not ask whether an instrument measures what it purports to measure, or whether it satisfies the axioms of measurement. It asks only whether an instrument is

familiar, widely used, and acceptable to stakeholders. In doing so, it confuses consensus with science. Consensus may establish a shared language, but it cannot transform numbers into measures. A multiattribute ordinal total remains ordinal no matter how many experts agree to endorse it. The result is that ICHOM's "standard sets" are not standards of measurement but conventions of convenience.

The consequences are profound. When ICHOM endorses an instrument, it signals to the global health community that the tool is fit for purpose. Researchers incorporate it into trials. Health systems embed it in registries. Policymakers cite it in evaluations. But if the tool does not generate measures, the downstream claims are numerology. A therapy may appear to improve a summed score by three points, but what does that mean? Without interval properties, the difference cannot be interpreted as a magnitude. Without invariance, the difference may not hold across subgroups. Without unidimensionality, the difference may reflect shifts in multiple unrelated attributes rather than a coherent change in one. In short, the claim cannot be evaluated as true or false, only repeated as a number with the appearance of precision.

The failure is not a matter of obscure technicalities. The standards of measurement were available almost 60 years before ICHOM began. Stevens had drawn the distinction between ordinal and interval data as early as 1946, warning that misuse of numbers outside their admissible transformations leads only to nonsense. Suppes in the 1950s advanced the axioms of extensive measurement, establishing the formal conditions under which empirical concatenations could sustain additive structure [6]. Luce and Tukey in the 1960s codified the axioms of additive conjoint measurement, specifying the cancellation conditions that guarantee meaningful numerical representation [7]. These strands were brought together definitively in 1971 with the publication of *Foundations of Measurement, Volume I*, which made clear the representational and uniqueness theorems on which all legitimate measurement must rest [3]. Rasch, already in 1960, had shown how ordinal responses could be transformed into interval scales under a probabilistic model that satisfied these axioms. And Wright in 1977 demonstrated that Rasch was not just another item response model, but the only one consistent with representational measurement theory [8]. By the time ICHOM was founded, these were not esoteric insights but established science for 40 years. Yet they were ignored. The ICHOM consortium pressed ahead with a program that bypassed measurement in favor of consensus.

This neglect raises the uncomfortable question of credibility. If ICHOM's instruments cannot sustain arithmetic, then the claims built upon them are not scientific claims. They are at best descriptions, at worst misleading numbers. To present them as measures is to mislead stakeholders into believing that outcomes are being rigorously assessed when in fact they are not. It is to cultivate an illusion of precision without its reality.

ICHOM fails because it has confused the task of choosing outcomes with the task of establishing measures. Outcomes can only be meaningful if they rest on measurement. Without that foundation, every subsequent layer, comparisons, registries, value-based purchasing collapses. The consortium has succeeded in branding, convening, and standard setting, but it has failed in the one thing that matters: ensuring that what it endorses are measures. Unless it confronts this failure, ICHOM will not survive as a scientific enterprise. It will persist only as a clearinghouse for non-measures, perpetuating the very confusion it was meant to resolve.

PROTOCOLS AND CLAIMS

If ICHOM is to move beyond its current role as a catalogue of consensus instruments, it must recognize that measurement is not an accessory to claims but their foundation. A claim about therapy response has meaning only to the extent that it is supported by a protocol that specifies the attribute, the instrument, and the analytic framework. The protocol provides the warrant for the claim: it states what is being measured, how it is being measured, how invariance will be tested, and what thresholds will define clinically meaningful change. Without such a protocol, a claim is not a scientific proposition but a hope dressed in numbers.

The task of ICHOM should therefore be recast from assembling questionnaires to endorsing protocols. A protocol begins with a clear definition of the value claim. For manifest attributes the claim might be that a therapy reduces hospital days, increases time in range, or improves walking distance by a specified margin. For latent attributes the claim might be that a therapy reduces dyspnea severity, depressive symptom burden, or fatigue interference, each defined as a single construct. The next step is to declare the scale type that underwrites the claim. Manifest attributes must be measured on ratio scales with true zeros and units that permit proportional comparisons. Latent attributes must be captured by Rasch-calibrated logit scales, ensuring unidimensionality, ordered thresholds, invariance across populations, and interval spacing. Only when the scale type is explicit can the claim be said to rest on admissible measurement.

The protocol must also state the expected timeframe of evaluation. Claims are not timeless; they must be evaluated over intervals that are clinically relevant and practically observable. A protocol might stipulate twelve months for compliance claims, six months for functional improvement, or three months for symptom reduction. The timeframe anchors the claim in empirical reality and makes it falsifiable: the therapy either achieves the specified change within the defined interval or it does not. Equally essential is the definition of minimally important differences. For linear ratios this may be expressed in days, metres, or units; for Rasch scales it must be expressed in logits, supported by evidence of interpretability and reproducibility. Without these thresholds, claims risk collapsing into vague promises rather than testable propositions.

What follows from this orientation is that ICHOM should cease to promote “standard sets” of mixed instruments and instead promote libraries of protocols tied to admissible measures. Each disease area would be defined not by a bundle of questionnaires but by a portfolio of value claims, each with its own protocol. These protocols would specify the measure, the timeframe, the threshold for meaningful change, and the procedures for ensuring invariance. A claim to reduce fatigue severity, for example, would be supported by a Rasch-calibrated fatigue scale with documented fit statistics, DIF analysis across age and gender, and an explicit minimally important difference in logits. A claim to reduce hospitalizations would be supported by a protocol specifying the count method, censoring rules, and time horizon. In both cases the claim is meaningful because the measurement is defensible.

The importance of this shift cannot be overstated. The entire rationale of ICHOM is to provide a platform for international comparability and benchmarking. Yet comparability cannot be achieved by consensus alone; it must be built into the structure of the measures themselves. Only ratio scales and Rasch-calibrated logit scales permit the arithmetic of comparison, the pooling of results across

settings, and the evaluation of claims as true or false. Protocols enforce this discipline. They ensure that every claim is tied to a measure, every measure to a scale type, and every evaluation to a reproducible standard.

The survival of ICHOM therefore depends on its willingness to pivot from instruments to protocols, from consensus to science. Protocols anchored in RMT are not optional; they are the only path to credible, evaluable, and replicable claims. Without them, ICHOM remains a warehouse of non-measures. With them, it becomes the steward of measurement and the guarantor of outcomes that can be trusted.

Illustration: heart failure and dyspnea claim

Consider the claim that a new therapy reduces dyspnea severity in patients with chronic heart failure over a six-month period. The protocol must first define the attribute. Dyspnea is not a composite of fatigue, exercise tolerance, and mood; it is a single latent construct reflecting the subjective experience of breathlessness. The attribute is latent, which means it cannot be directly observed but must be inferred from patient responses to carefully designed items.

The measure must therefore be a Rasch-calibrated logit scale. Items are written to reflect ordered gradations of breathlessness, each with response categories that are monotonic and unidimensional. Pilot testing ensures ordered thresholds, fit statistics within acceptable ranges, and the absence of local dependence. Differential item functioning is tested across sex, age, and language groups, and items showing bias are revised or removed. The resulting instrument yields person measures in logits, centered on the sample mean, and capable of being linked across versions through anchor items.

The claim is then stated explicitly: patients receiving therapy will demonstrate a mean reduction of at least 0.5 logits on the dyspnea scale at six months compared to baseline. This threshold is justified by prior validation work showing that 0.5 logits corresponds to a minimally important difference detectable by patients and associated with observable improvements in physical functioning. The protocol specifies the timeframe (six months), the unit of measurement (logits), the analytic approach (Rasch analysis with fit evaluation), and the falsification criterion (failure to achieve the 0.5 logit reduction). This claim can be replicated, falsified, and compared across health systems because it is anchored in a true measurement structure.

Illustration: heart failure and days alive out of hospital

Now consider a claim for the same therapy: that it increases the number of days patients are alive and out of hospital during the first twelve months after initiation. This attribute is manifest, not latent. It is directly observable in administrative or clinical records, expressed as a count of days. By definition, it is a linear ratio scale with a true zero (no days) and equal intervals (each day is of the same length).

The claim is then straightforward: patients on the therapy will average at least 30 more days alive and out of hospital over twelve months compared to patients on standard care. The protocol sets the timeframe (twelve months), defines the unit (days), and specifies data sources (hospital

discharge records, mortality data). Because the measure is manifest and ratio-scaled, it requires no transformation: the arithmetic is legitimate, the differences meaningful, and the claim directly falsifiable.

LATENT CONSTRUCTS, TRAITS AND POSSESSION

To understand what it means to measure health outcomes rigorously, we must return to the notion of the latent construct. A latent construct is an attribute that cannot be observed directly but is inferred through its manifestations. In every disease area there are constructs of this kind: fatigue in cancer, dyspnea in heart failure, depression in mental health, treatment burden in diabetes. These are not tangible, countable phenomena like days alive or units of insulin dispensed. Rather, they are experiential states that patients report, often through ordered response categories on a questionnaire. To say that a therapy reduces fatigue or lessens dyspnea is to make a claim about a latent construct. But because these constructs are not directly observable, measurement requires an explicit model that links what patients report to an underlying continuum. Without such a model, one is left with nothing more than ordinal labels and summed scores that cannot sustain the arithmetic of science.

The crucial distinction here is between the construct itself and the manifestations or traits that compose it. Consider the case of heart failure. Breathlessness is one of its most characteristic symptoms, but it is not the only dimension of patient experience. There may also be fatigue, fluid retention, limitation of social participation, or diminished physical capacity. Each of these is a distinct trait embedded within the broader construct of living with heart failure. The task of measurement is not to collapse all these manifestations into a single number, but to isolate each trait and develop a measure that captures variation along its continuum. This is why Rasch measurement is indispensable: it demands unidimensionality. If the set of items in an instrument does not reflect a single underlying trait, the model will reject them. By contrast, summed score instruments like the Kansas City Cardiomyopathy Questionnaire (KCCQ) bundle together multiple traits and then treat the total as though it reflected one continuum (see Part 1). This is precisely what RMT forbids.

Focusing on a single manifestation or trait makes the claim precise and testable. Suppose the claim is that a therapy reduces dyspnea severity. The trait of interest is the subjective experience of breathlessness, not fatigue, not social participation, and not general quality of life. To measure this, items are written to capture gradations in the experience of breathlessness: perhaps shortness of breath when climbing stairs, when dressing, when walking across a room. Responses are given on ordered categories. These responses are then calibrated with the Rasch model. Each item is placed on a logit scale of difficulty, and each patient is placed on the same logit scale of ability or possession. The number line is the bridge between the raw categorical data and interval measurement. At baseline, each patient has a possession score in logits that tells us where they stand on the continuum of dyspnea severity. This is the starting point for any evaluation of therapy impact.

What is distinctive about the Rasch logit scale is that it represents constant relative differences. Moving one logit higher on the scale represents the same proportional increase in the odds of endorsing a more severe category, no matter where one is on the continuum. This property is what

makes the scale interval, and it is what makes possession of a trait measurable. At baseline, a patient may have a possession of dyspnea severity at 1.2 logits. After six months on therapy, the same patient may record a possession of 0.6 logits. The difference is 0.6 logits, which has a precise meaning: it reflects a proportional change in the likelihood of reporting more severe categories of breathlessness. Aggregated across patients, the average logit change can be analyzed statistically. Effect sizes can be computed, comparisons to a comparator therapy can be made, and the claim can be confirmed or falsified. Unlike ordinal sums, which cannot support meaningful differences, the logit scale preserves structure and allows arithmetic.

Some readers may be more comfortable with scales that range from 0 to 100. Rasch allows this, but only through linear transformation. The underlying metric remains the logit, but for presentation purposes it can be rescaled. A logit distribution spanning –3 to +3 can be linearly mapped onto a 0 to 100 scale, preserving interval properties while providing a more familiar frame for interpretation. It must be emphasized, however, that this is a cosmetic change: the arithmetic and statistical analysis remain in logits. To transform logits into percentages is to misrepresent them, because percentage implies ratio scaling with a true zero and meaningful doubling. Logits are interval, not ratio, and must be respected as such.

That said, Rasch also allows the interval logit scale to be transformed into a ratio form under particular conditions. If the logit continuum can be anchored with a true zero point that corresponds to the absence of the attribute, then proportional comparisons become meaningful. In practice this is rare, because most subjective attributes do not admit a natural zero. Breathlessness cannot be said to disappear entirely, nor can fatigue or distress. But in cases where a true zero is definable, a ratio transformation is possible. Otherwise, the scale remains interval, which is already a substantial advance over ordinal sums.

The significance of possession is that it makes therapy impact evaluable. When we say that a therapy reduces dyspnea severity by 0.6 logits compared to a comparator, we are making a claim that can be falsified. It is not a narrative about multiattribute quality of life, a composite score of multiple dimensions, nor a preference index. It is a precise statement about movement along a unidimensional continuum, expressed in an interval metric, supported by Rasch calibration. This is the foundation of science: claims that can be tested against data.

The alternative, which remains the default in most disease areas, is pseudo-measurement. Questionnaires are bundled, scores summed, and changes reported as if they were meaningful. But a change of 6 points on a summed score has no defined meaning unless the scale is shown to be interval and unidimensional. Without Rasch calibration, such changes are arbitrary and cannot support evaluable claims. ICHOM's current standard sets fall into this trap: they endorse questionnaires without ensuring that the traits are unidimensional or that the scales meet the axioms of measurement.

By contrast, Rasch re-engineering focuses on traits, calibrates them onto logit scales, and produces possession profiles that can be tracked over time. A therapy's impact is then a shift in possession, which can be compared across groups, tested for statistical significance, and expressed in a form that respects measurement theory. This is how latent constructs become measurable and how therapy claims become scientific.

In short, latent constructs must be decomposed into traits, traits must be modeled with Rasch, and possession must be the language of therapy impact. This approach makes it possible to replace narrative storytelling with scientific evaluation. It is the only way to bring latent constructs into the domain of credible, evaluable, and replicable health technology assessment.

MEETING REPRESENTATIONAL AXIOM REQUIREMENTS

It must be emphasized at the outset that the axioms of RMT apply with equal force to both manifest and latent attributes. The distinction lies not in whether the axioms matter, but in how they are satisfied. Manifest attributes, such as time, weight, or blood glucose concentration, are directly observable and lend themselves to empirical verification. Latent attributes, such as pain interference or dyspnea severity, are not directly observable and must be inferred from structured responses. In both cases, the standards of measurement are the same: claims can only rest on numbers that preserve empirical structure and admit the operations of arithmetic defined by the relevant scale type. Anything less is pseudo-measurement; however useful it may appear for administrative purposes.

For manifest attributes the task is relatively straightforward. The analyst must apply a checklist to ensure that the proposed attribute meets the standards of a ratio scale. Does the attribute admit a true zero, such that absence can be defined? Are intervals between successive observations demonstrably equal? Can proportional comparisons be made, such that twice the quantity represents twice the empirical magnitude? Only when these conditions are met can one proceed to arithmetic operations, statistical analysis, and claims framed in scientific language. Where the conditions are not met, numbers may still be generated, but they cannot be treated as measures. A patient's weight in kilograms is a measure because it has a true zero, equal intervals, and invariance across observers and instruments. A blood pressure reading, when calibrated, likewise qualifies. But an ordinal rating of symptom frequency, scored from "never" to "always," cannot, without more, sustain ratio operations. The checklist is indispensable to guard against the incursion of non-measures into scientific claims.

Latent constructs present a more demanding challenge. Because they cannot be observed directly, the question becomes how to ensure that the numbers derived from item responses are legitimate measures. This is where the Rasch model provides the indispensable bridge. Unlike other statistical models, which seek only to fit data, Rasch begins with a probabilistic structure designed to embody the axioms of measurement. It asserts that the probability of a response depends solely on the difference between person ability (or possession of the latent trait) and item difficulty, expressed on the same logit scale. If the data fit the model, then the requirements of unidimensionality, additivity, and invariance are met. This is what makes Rasch unique: it operationalizes the axioms in a way that data can confirm or reject.

The singular contribution of Wright in 1977 was to make this connection explicit. He showed that the Rasch probabilistic framework did not merely produce useful scores but instantiated the requirements of representational measurement. By demonstrating that the Rasch model yields conjoint simultaneous measurement of persons and items on a common interval scale, Wright closed the gap between abstract axioms and applied practice. What had seemed an insurmountable problem, deriving interval measures from categorical responses, was given a rigorous solution. If

the data fit the Rasch model, then the resulting logit scale preserves the structure required by the axioms. If the data do not fit, then the instrument fails and must be revised. There is no middle ground.

The difference, then, is not in the demands of the axioms but in the method of their realization. For manifest attributes, the checklist confirms that empirical observations already satisfy the requirements of a ratio scale. For latent attributes, the Rasch model ensures that the responses can be transformed into an interval scale, provided the data conform to the model. In both cases, measurement is the arbiter, and only by meeting these requirements can health technology assessment claim to operate within the domain of science.

CONCLUSION: SUCCESS OR FAILURE

The task confronting ICHOM is interesting; it determines whether there is a renunciation of adherence to failed measures, or a decision to circle the wagons and maintain the measureless *status quo*. From the perspective of RMT and falsification the choice is obvious, but demanding. To continue endorsing disease-specific questionnaires without a filter grounded in measurement theory is to remain in the realm of pseudo-science, perpetuating instruments that yield ordinal scores, collapse multiple attributes into spurious totals, and deny unidimensionality. Such instruments may generate numbers, but they do not generate measures; only ridicule. To sustain them is to trade rigor for convenience and to present health systems with evidence that cannot be falsified.

It is not as though this is a new challenge. The requirements for measurement have been debated and settled for decades. The axioms of representational measurement were clarified and codified by the early 1970s. By that time Stevens had drawn the line between numbers and measures, Suppes had set out the axioms of extensive measurement, Luce and Tukey had demonstrated cancellation, and the first volume of *Foundations of Measurement* had provided the definitive synthesis. Rasch had already demonstrated how ordinal responses could be transformed into interval measures, and Wright in 1977 made explicit Rasch's unique alignment with the axioms. For the 80 years since Stevens seminal contribution these standards have been available, while HTA and ICHOM have chosen to ignore them. The result has been the embrace by HTA practitioners of false measurement claims masquerading as science.

The alternative is clear, but requires courage. ICHOM can define itself as the steward of measurement in health outcomes by adopting a charter that no claim will be endorsed unless it rests on either a linear ratio scale for manifest attributes or a Rasch logit ratio scale for latent constructs. This would mean abandoning the comfort of consensus instruments in favor of a systematic rebuilding. It would mean confronting the reality that many widely used tools must be retired or re-engineered. It would mean educating stakeholders in the logic of Rasch measurement and insisting that only unidimensional constructs, calibrated on invariant scales, can sustain claims of therapy impact.

Success will not be measured by the size of ICHOM's catalogue but by the credibility of its measures. If the organization takes this path, it can become the global benchmark for outcomes measurement, reshaping not just how diseases are evaluated but how therapies are judged,

reimbursed, and improved. Failure will mean further entrenchment in relativism, producing numbers that masquerade as measures but collapse under the scrutiny of science. The decision is stark: either a pivot to the discipline of representational measurement and falsification or a continued allegiance to instruments that deny them. ICHOM's survival depends on choosing science over consensus.

ACKNOWLEDGMENT

Portions of this paper were drafted and edited with assistance from ChatGPT (version 5; OpenAI). The author reviewed, verified, and refined all AI-assisted text and assumes full responsibility for the accuracy, integrity, and originality of the final content.

REFERENCES

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

[2] Popper K. *Objective Knowledge: An Evolutionary Approach*. Revised edition. Oxford: Clarendon Press, 1979

[3] Krantz D, Luce R, Suppes P, Tversky A**.** *Foundations of Measurement*, Volumes I–III. New York: Academic Press, 1971 (Vol. I); 1989 (Vol. II); 1990 (Vol. III)

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

[5] Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded ed., Chicago: University of Chicago Press, 1980)

[6] Suppes P. A Set of Independent Axioms for Extensive Quantities. *Portugaliae Mathematica* 1951; 10: 163–172

[7] Luce R, Tukey J. Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. J *Math Psychol. 1964;* 1(1): 1–27

[8] Wright B. "Solving Measurement Problems with the Rasch Model." *Journal of Educational Measurement* 14, no. 2 (1977): 97–116