AI LLM TRUE or FALSE

Statement: *"Interval measures lack a true zero."*
Classification: TRUE

An interval scale, by definition, does not possess a true zero. This is not a matter of convention or preference; it follows directly from the axioms of representational measurement theory. A true zero is not simply the lowest observed value on a scale. It is a point that represents the complete absence of the attribute being measured. For a scale to have a true zero, zero must be meaningful in an absolute sense, such that ratios involving zero and nonzero values have interpretable meaning. Interval scales do not satisfy this requirement.

The defining property of an interval scale is that equal numerical differences represent equal differences in the underlying attribute. This permits addition and subtraction, but it does not permit multiplication or division in a meaningful way. The location of zero on an interval scale is arbitrary: it can be shifted without altering the empirical meaning of differences. Temperature measured in Celsius provides the canonical example. Zero degrees Celsius does not represent the absence of temperature; it represents an arbitrarily chosen point on the temperature continuum. The same physical state can be expressed as 0°C, 32°F, or 273.15 K, depending on the scale origin. Because the zero point can be relocated by a linear transformation without changing the meaning of the measure, it cannot represent absence of the attribute.

This arbitrariness of the zero point is precisely what distinguishes interval scales from ratio scales. In a ratio scale, zero is fixed by the empirical structure of the attribute. Length, mass, and time have true zeros: zero length means no length, zero mass means no mass, zero time means no duration. Because zero represents absence, ratios are meaningful. An object that is two meters long is twice as long as an object that is one meter long. Such statements are meaningless on an interval scale, where the zero does not anchor the scale to absence.

The absence of a true zero has direct implications for permissible arithmetic. On an interval scale, statements such as "twice as much" or "half as much" are invalid, because multiplication and division depend on a meaningful zero. Only differences are interpretable. Saying that one temperature is 10 degrees higher than another is meaningful; saying it is twice as hot is not. This limitation is not a technical inconvenience; it is a categorical boundary that protects arithmetic from misrepresentation.

In the context of health technology assessment, this distinction is routinely ignored. Preference-based utility scores are often treated as if they were ratio measures, even when they at best satisfy interval properties, and often not even that. By treating interval-scale quantities as if they possessed true zeros, HTA practice enables illegitimate arithmetic operations, including multiplication by time and division by cost. Recognizing that interval measures lack a true zero is therefore not a pedantic observation. It is a foundational constraint. Once it is acknowledged, large parts of standard HTA arithmetic become impossible, not controversial.

Statement: *"Measures must be unidimensional."*
Classification: TRUE

Unidimensionality is a necessary condition for measurement. A measure can represent only one attribute at a time. This requirement is not a stylistic preference or a simplifying assumption; it follows directly from the logic of representational measurement theory. For numbers to represent an empirical attribute in a meaningful way, there must be a single, well-defined dimension of variation that those numbers correspond to. If more than one attribute is involved, the numerical representation becomes ambiguous, and arithmetic loses interpretability.

The core purpose of measurement is to map variations in a single empirical attribute onto variations in numbers while preserving relevant relations. If multiple attributes are conflated, there is no longer a unique empirical structure being represented. A numerical difference could reflect a change in one attribute, another, or some mixture of both. In such cases, the numbers do not correspond to anything determinate in the real world. They become labels attached to heterogeneous bundles rather than measures of a single property.

This is why unidimensionality is foundational in all mature measurement sciences. Length, mass, time, temperature, and electric charge are each defined along a single dimension. When phenomena are multidimensional, science does not respond by inventing a single composite measure and pretending it is one thing. Instead, it measures each dimension separately. Velocity is not measured directly; it is derived from two unidimensional measures, distance and time, each with its own scale properties. The derivation is lawful precisely because the underlying measures are unidimensional and their scale types are known.

In psychometrics and the measurement of latent traits, unidimensionality is equally essential. Models such as the Rasch model exist precisely to test and enforce unidimensionality. Rasch measurement does not assume that a set of items measures a single trait; it evaluates whether the data conform to that requirement. Only if responses can be explained by variation along one latent dimension can a scale be constructed. Without unidimensionality, there is no latent trait to measure, only a collection of loosely related indicators.

The consequences of ignoring unidimensionality are severe. Composite indices that combine multiple attributes into a single score cannot be interpreted as measures because changes in the score do not correspond to changes in any single attribute. Arithmetic performed on such composites is uninterpretable, regardless of how sophisticated the weighting scheme appears. The problem is not that the weights are debatable; it is that no weighting can rescue the loss of dimensional coherence.

In health technology assessment, this requirement is routinely violated. Constructs such as "health-related quality of life" are treated as if they were single attributes, even though they explicitly combine distinct dimensions such as mobility, pain, mood, and self-care. When such multidimensional constructs are collapsed into a single index, the result is not a measure of anything. It is a summary score. Treating that score as if it were a measure enables arithmetic that has no empirical meaning.

To insist that measures must be unidimensional is therefore not to deny the complexity of health or human experience. It is to insist that complexity be respected rather than obscured. Measurement requires discipline. Without unidimensionality, numbers do not measure; they merely summarize.

Statement: *"Multiplication requires a ratio measure."*
Classification: TRUE

Multiplication is only meaningful when applied to quantities that possess ratio-scale properties. This is not a convention of statistics or an assumption of modeling practice; it is a logical requirement of arithmetic grounded in representational measurement theory. A ratio scale is defined by two essential properties: equal intervals and a true zero. Without both, multiplication and division cannot preserve empirical meaning.

The role of a true zero is decisive. A true zero represents the complete absence of the attribute being measured. It anchors the scale in the empirical world, making ratios interpretable. When a quantity has a true zero, statements such as "twice as much," "half as much," or "three times larger" have meaning because zero fixes the origin of the scale in a non-arbitrary way. Length, mass, time, and count all satisfy this condition. Zero length means no length, zero mass means no mass, and zero time means no duration. Because of this, multiplication on these quantities corresponds to real-world relations.

Interval scales do not satisfy this requirement. Although they permit addition and subtraction, their zero point is arbitrary and can be shifted without changing the meaning of differences. Because the origin is not fixed by the attribute itself, ratios are meaningless. Saying that 20 degrees is twice as hot as 10 degrees has no physical interpretation if temperature is measured on an interval scale such as Celsius or Fahrenheit. The numerical ratio does not correspond to any ratio in the underlying attribute. Multiplication in this context produces a number, but not a meaningful quantity.

This distinction is categorical, not gradual. A scale either has a true zero or it does not. If it does not, multiplication is forbidden. There is no mathematical workaround, no weighting scheme, and no modeling sophistication that can overcome this constraint. Performing multiplication on a non-ratio scale does not produce an approximate result; it produces nonsense. The arithmetic operation ceases to represent anything empirical.

In measurement science, derived quantities are constructed through multiplication only when the scale types of the components justify it. Velocity is distance divided by time, both ratio measures. Force is mass multiplied by acceleration, again ratio measures. These constructions are lawful because the operands satisfy the axioms required for multiplication. If one operand lacked ratio properties, the derived quantity would be uninterpretable.

In health technology assessment, this rule is systematically violated. Preference-based utility scores, which at best can claim interval properties and often not even that, are multiplied by time to generate QALYs. This multiplication is treated as if it were analogous to multiplying meters by meters or seconds by seconds. It is not. Because utilities lack a true zero, the product has no

interpretable meaning. Time remains a ratio measure; utility does not. Multiplying a ratio measure by a non-ratio measure does not upgrade the latter. It contaminates the former.

The insistence that multiplication requires a ratio measure is therefore not pedantic. It is protective. It marks the boundary between arithmetic that represents reality and arithmetic that merely produces numbers. Once this boundary is crossed, the results cannot be defended as measures, no matter how widely they are used or how institutionally embedded they have become.

Statement: *"Time trade-off preferences are unidimensional."* **Classification: FALSE**

Time trade-off (TTO) preferences are not unidimensional, and treating them as such is a fundamental error. Unidimensionality requires that responses vary along a single underlying attribute, such that differences in observed values can be attributed solely to differences in that one dimension. TTO data fail this requirement because they simultaneously reflect multiple, conceptually distinct attributes that cannot be disentangled into a single latent dimension.

In a TTO task, respondents are asked to trade length of life against a described health state. The resulting preference value is therefore not a manifestation of a single attribute, but an amalgam of several. At a minimum, TTO responses reflect attitudes toward longevity, attitudes toward the quality of the described health state, attitudes toward death, risk perception, time preference, loss aversion, and task comprehension. None of these components is separable within the elicited value. A change in a TTO score cannot be uniquely attributed to a change in "health quality," because it may equally reflect a change in willingness to sacrifice life years, fear of death, or discounting of future time.

This multidimensionality is structural, not incidental. It arises from the very design of the TTO task. By construction, the respondent is asked to consider two qualitatively different attributes, time and health state, and to make a judgment that balances them. The output is therefore a trade-off function, not a measure of a single trait. Even if all respondents perfectly understood the task and responded consistently, the resulting numbers would still represent a compound preference relation rather than variation along one dimension.

Attempts to treat TTO values as unidimensional often rest on a category error: confusing a single numerical output with a single underlying attribute. Producing one number does not guarantee unidimensionality. A composite index can always be expressed as a scalar, but that scalar does not correspond to a single empirical dimension unless the contributing attributes are demonstrably aligned. In TTO, there is no empirical or theoretical basis for claiming such alignment. Indeed, empirical evidence routinely shows that TTO responses vary systematically with factors unrelated to health state severity, such as age, framing, time horizon, and cultural attitudes toward death.

From a measurement perspective, unidimensionality is not something that can be assumed; it must be tested. In latent trait measurement, models such as Rasch explicitly evaluate whether responses conform to a single latent dimension. TTO methods provide no such test. They simply

assume that the elicited preference reflects a single quantity called “utility,” even though the task embeds multiple dimensions by design.

In health technology assessment, treating TTO preferences as unidimensional enables further illegitimate steps, including treating the resulting scores as interval or ratio measures and multiplying them by time. Once unidimensionality fails, these downstream operations lose any claim to meaning. The falsity of the statement that TTO preferences are unidimensional is therefore not a minor technical point. It exposes a foundational flaw: TTO outputs are not measures of a single attribute at all, but context-dependent preference constructions masquerading as quantities.

Statement: *“Ratio measures can have negative values.”*
Classification: FALSE

Ratio measures cannot have negative values, and this follows directly from the defining properties of a ratio scale. A ratio scale is characterized by two essential features: equal intervals and a true zero. The true zero is not a conventional reference point; it represents the complete absence of the attribute being measured. Because zero denotes absence, values on a ratio scale are bounded below by zero. Negative values are therefore conceptually and mathematically impossible.

The presence of a true zero is what distinguishes ratio scales from interval scales. On a ratio scale, zero is fixed by the empirical structure of the attribute itself. Zero length means no length, zero mass means no mass, zero duration means no time, and zero count means nothing is present. Once the absence of the attribute is reached, there is no further decrement possible. The scale cannot extend below zero without violating its empirical meaning. Negative length, negative mass, or negative time do not represent lesser amounts of the attribute; they represent conceptual contradictions.

This constraint is not relaxed by mathematical convenience or modeling practice. While mathematics allows negative numbers in the abstract, representational measurement theory restricts their use to situations where they correspond to meaningful empirical relations. For ratio scales, negative numbers do not correspond to any possible state of the attribute. As a result, they are excluded by definition. Allowing negative values would destroy the interpretability of ratios, because ratios rely on zero as a meaningful origin. If negative values were permitted, statements such as “twice as much” or “half as much” would lose coherence.

Interval scales, by contrast, can and often do include negative values precisely because they lack a true zero. Temperature measured in Celsius or Fahrenheit can take negative values because zero does not represent the absence of temperature; it represents an arbitrary point on the scale. Shifting the zero point does not change the meaning of differences, which is why interval scales permit both positive and negative numbers. This is exactly what ratio scales do not permit. Their zero point is fixed and non-arbitrary, which precludes negative values.

In health technology assessment, the confusion between interval and ratio scales is pervasive, and it leads directly to the mistaken acceptance of negative “ratio” values. Utility scores derived

from preference elicitation methods are sometimes allowed to fall below zero and are then treated as if they were ratio measures. This is internally inconsistent. A quantity that admits negative values cannot, by definition, be a ratio measure, because the presence of negative values signals the absence of a true zero.

The claim that ratio measures can have negative values is therefore false in a categorical sense. It is not a matter of degree or approximation. A scale that permits negative values has already abandoned the defining property that makes ratio arithmetic meaningful. Recognizing this boundary is essential, because once negative values are admitted, any subsequent multiplication, division, or ratio comparison becomes uninterpretable.

Statement: *"EQ-5D-3L preference algorithms create interval measures."*
Classification: FALSE

EQ-5D-3L preference algorithms do not create interval measures. They generate preference scores that at best preserve ordinal information and, in many cases, do not even satisfy the minimal requirements for interval scaling. Treating these outputs as interval measures is a category error that arises from confusing numerical estimation with measurement.

An interval scale requires more than numerical spacing. It requires that equal numerical differences correspond to equal differences in the underlying attribute across the entire scale. This property must be justified empirically, not assumed. Interval measures also require invariance: the meaning of a unit difference must be independent of the particular items, respondents, or estimation sample used. EQ-5D-3L preference algorithms satisfy none of these conditions.

The EQ-5D-3L descriptive system is explicitly multidimensional, combining mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Preference algorithms then assign weights to levels within each dimension based on population elicitation exercises, typically using time trade-off or related methods. The resulting index value is a weighted sum of responses across dimensions. Producing a single number does not create an interval measure. It merely produces a composite score. Without evidence that the weighted combination represents variation along a single underlying attribute with equal intervals, interval status cannot be claimed.

Crucially, the algorithms impose interval structure rather than discovering it. The spacing between health states is determined by regression coefficients estimated from preference data that themselves lack interval properties. The model outputs are therefore artifacts of the chosen functional form, anchoring rules, and sample characteristics. A different elicitation method, valuation protocol, or population produces a different set of coefficients and a different numerical scale. This dependence on estimation context violates the invariance requirement of interval measurement.

The problem is compounded by the properties of the underlying preference data. Time trade-off responses do not generate interval-scale observations, as they do not satisfy unidimensionality or possess a true zero. Aggregating and modeling such data cannot magically upgrade their scale properties. Statistical estimation does not transform ordinal or mixed-attribute preferences into interval measures. It only fits numbers to responses.

Contrast this with legitimate interval measurement in the human sciences, which requires explicit testing of scale properties, often through models such as Rasch that enforce unidimensionality and invariant item calibration. EQ-5D-3L algorithms do not test whether equal differences in index scores correspond to equal differences in any empirical attribute. They assume it, because the downstream arithmetic of QALYs requires it.

In health technology assessment, labeling EQ-5D-3L preference scores as interval measures enables subtraction, averaging, and multiplication that would otherwise be impermissible. But this labeling is rhetorical, not scientific. Without demonstrated equal-interval properties and invariance, EQ-5D-3L preference algorithms do not create interval measures. They create numerically convenient indices, and the distinction matters because arithmetic legitimacy depends on it.

Statement: *"The QALY is a ratio measure."*
Classification: FALSE

The QALY is not a ratio measure, and it fails the defining requirements of ratio measurement in multiple, independent ways. A ratio measure must possess equal intervals and a true zero, and it must represent variation along a single, well-defined attribute. The QALY satisfies none of these conditions. Its continued treatment as a ratio measure in health technology assessment reflects institutional convention rather than measurement logic.

The QALY is constructed by multiplying time, which is a genuine ratio measure, by a utility score intended to represent health-related quality of life. The scale properties of the product cannot exceed the weakest component. Even if time is ratio-scaled, the utility component is not. Preference-based utility scores lack a true zero, are not demonstrably interval-scaled, and do not represent a single unidimensional attribute. Multiplying time by such a score does not create a ratio measure; it contaminates a ratio quantity with a non-measure.

The absence of a true zero is decisive. For the QALY to be a ratio measure, zero QALYs would have to represent the complete absence of the attribute being measured. In practice, zero QALYs corresponds to zero time, not zero "health." The utility component does not have a true zero, as its zero point is arbitrarily defined through anchoring conventions such as "dead = 0." These anchors are not empirically grounded absences of health, and they can vary across valuation protocols and populations. A quantity whose zero depends on convention rather than absence cannot support ratio interpretation.

The QALY also fails unidimensionality. It purports to represent a single quantity of "health," but in fact conflates duration with a multidimensional, preference-weighted index of health states. Duration and health state quality are distinct attributes. Combining them does not yield a single

underlying dimension; it yields a composite. Composite quantities cannot be ratio measures because changes in the composite do not correspond to changes in any single attribute.

The existence of negative QALYs further exposes the category error. In some valuation systems, utility scores are allowed to fall below zero, implying "states worse than dead." When such values are multiplied by time, the result is a negative QALY. A ratio measure cannot take negative values, because zero represents absence of the attribute. The mere possibility of negative QALYs is sufficient to refute the claim that QALYs are ratio-scaled.

Finally, ratio measures permit meaningful ratio statements. If the QALY were a ratio measure, it would be meaningful to say that one intervention produces twice as much health as another. Such statements are routinely implied in cost-per-QALY calculations, but they are indefensible. Because the underlying utility differences are not equal-interval and lack a true zero, the ratios of QALYs do not correspond to ratios of any empirical attribute.

The claim that the QALY is a ratio measure is therefore false in a categorical sense. It is not approximately false or philosophically questionable; it is mathematically impossible. The QALY is a constructed index whose numerical properties are assumed for convenience. Treating it as a ratio measure enables arithmetic that has no empirical meaning, and once that is recognized, the central quantitative pillar of cost-effectiveness analysis collapses.

Statement: *"Time is a ratio measure."*
Classification: TRUE

Time is a ratio measure because it satisfies all the defining requirements of ratio-scale measurement. It possesses equal intervals, a true zero, and invariant units, and it supports meaningful multiplication, division, and ratio comparisons. These properties are not matters of convention within health technology assessment; they are grounded in the empirical structure of time itself and have been recognized across the physical sciences for centuries.

The defining feature of a ratio scale is the existence of a true zero that represents the complete absence of the attribute being measured. For time, zero denotes no duration. A time interval of zero seconds means that no time has elapsed. This is not an arbitrary reference point that can be shifted without consequence. It is fixed by the empirical meaning of duration. There is no meaningful sense in which a negative duration exists in the empirical world. This fixed zero anchors the scale and allows ratios to be interpreted.

Time also has equal intervals. One second represents the same duration regardless of when it occurs or what is being timed. The difference between one and two seconds is empirically equivalent to the difference between ten and eleven seconds. This invariance of unit size is essential for arithmetic. It ensures that addition and subtraction correspond to the concatenation or removal of equal durations. Without equal intervals, even simple summation would be meaningless.

Because time has a true zero and equal intervals, multiplication and division are meaningful. It is coherent to say that one event lasted twice as long as another, or that an intervention extended

life by three times the duration achieved by a comparator. Ratios of time correspond to ratios in the underlying attribute. This is exactly what ratio-scale measurement entails. Time therefore supports all standard arithmetic operations without violating measurement axioms.

The ratio-scale properties of time are preserved across different units of measurement. Seconds, minutes, hours, and years are related by constant multiplicative transformations. Changing units rescales the numbers but does not alter their ratios. An event that lasts two hours is twice as long as one that lasts one hour, just as an event that lasts 120 minutes is twice as long as one that lasts 60 minutes. This invariance under multiplicative transformation is a hallmark of ratio scales.

In health technology assessment, time is often invoked correctly as a ratio measure, particularly when discussing survival, duration of treatment, or length of follow-up. Problems arise not because time lacks ratio properties, but because time is combined with quantities that do not share those properties. Multiplying time by non-ratio measures does not confer ratio status on the product. The ratio nature of time is not contagious; it cannot rescue illegitimate arithmetic.

Affirming that time is a ratio measure is therefore uncontroversial and foundational. It underscores the asymmetry at the heart of HTA arithmetic: one component of the QALY construction is lawful, the other is not. Recognizing this distinction is essential, because it clarifies that the problem lies not with the use of time, but with what is done to it.

Statement: *"Measurement precedes arithmetic."*
Classification: TRUE

Measurement must precede arithmetic because arithmetic operations are meaningful only when applied to quantities whose measurement properties are already established. This is a foundational principle of representational measurement theory and of mathematics as it is applied to the empirical world. Numbers do not acquire meaning simply by being manipulated; their meaning derives from the prior mapping between numbers and attributes. Without that mapping, arithmetic is not analysis but symbol manipulation detached from reality.

Measurement answers a logically prior question: *what kind of thing is being represented?* Before any arithmetic can be performed, it must be known whether the attribute admits ordering, equal intervals, or a true zero. These properties determine which operations are permissible. Ordinal scales permit ranking but not addition. Interval scales permit addition and subtraction but not multiplication. Ratio scales permit the full range of arithmetic operations. Arithmetic does not determine scale type; scale type determines arithmetic. Reversing this order destroys interpretability.

In all mature sciences, this ordering is taken for granted. Physicists do not multiply quantities until they have established what is being measured and on what scale. Engineers do not divide by variables whose dimensional properties are unknown. Units analysis, dimensional consistency, and scale properties are enforced before computation, not retrofitted afterward. Arithmetic is subordinate to measurement, not the other way around.

The error in health technology assessment is precisely the inversion of this logic. HTA begins with desired arithmetic outputs—cost-effectiveness ratios, aggregated QALYs, lifetime model results—and then assigns scale properties to inputs as needed to justify those operations. Utilities are treated as interval or ratio measures because the arithmetic requires them to be so, not because their measurement properties have been demonstrated. This is not measurement; it is rationalization.

Statistical estimation does not repair this inversion. Regression coefficients, preference weights, or model parameters do not create measurement properties. They merely produce numbers that conform to a chosen functional form. Without prior demonstration that the underlying variables possess the scale properties required for the intended arithmetic, the results remain uninterpretable. Arithmetic performed first and justified later cannot recover meaning.

The principle that measurement precedes arithmetic also explains why sensitivity analysis and uncertainty ranges cannot rescue invalid models. Varying inputs across plausible ranges does not address whether the operations themselves are lawful. One cannot test the robustness of an illegitimate multiplication. If the quantities are not measures, no amount of arithmetic refinement can make the results meaningful.

Affirming that measurement precedes arithmetic is therefore not methodological pedantry. It is the condition that separates science from numerology. Once this ordering is violated, numbers cease to represent attributes and become mere artifacts of calculation. Recognizing this principle forces a reckoning in HTA: many familiar calculations are not slightly flawed, but logically prohibited. Arithmetic can only follow measurement. When it leads instead, the result is not evidence, but illusion.

Statement: *"Summations of subjective instrument responses are ratio measures."*
Classification: FALSE

Summations of subjective responses are not ratio measures, and in most cases they are not measures at all. Adding together subjective responses—such as Likert-scale items, questionnaire scores, or patient-reported outcome responses—produces a numerical total, but it does not produce a quantity with ratio-scale properties. Treating such sums as ratio measures reflects a fundamental misunderstanding of both measurement theory and arithmetic.

A ratio measure requires three conditions: unidimensionality, equal intervals, and a true zero representing the complete absence of the attribute. Summated subjective scores fail on all three counts. First, unidimensionality is rarely demonstrated. Questionnaires typically include items that capture related but distinct aspects of experience, such as pain intensity, emotional distress, functional limitation, or satisfaction. Adding responses across items produces a composite score, not a measure of a single attribute. A single number does not imply a single dimension.

Second, subjective response categories do not have equal intervals. Likert-type scales (e.g., "no pain," "mild," "moderate," "severe") are ordinal. They indicate order but not magnitude. There is

no empirical basis for assuming that the difference between “mild” and “moderate” is equal to the difference between “moderate” and “severe.” Summing ordinal responses does not create interval structure; it merely compounds ordinal information. Arithmetic addition does not upgrade scale properties.

Third, summated subjective scores lack a true zero. A score of zero typically reflects a scoring convention, such as selecting the lowest response option on all items. It does not represent the complete absence of the underlying attribute. Zero pain on a questionnaire does not imply the absence of all pain-related experience; it implies endorsement of the lowest category as defined by the instrument. Because zero is not an empirical absence, ratio interpretations are impossible. Statements such as “twice as much pain” or “half the quality of life” have no meaning when based on such sums.

The fact that summated scores are widely used does not alter their measurement status. Statistical reliability, internal consistency, or responsiveness to change do not confer ratio properties. A score can be reliable and still not be a measure. Reliability concerns consistency of responses, not the legitimacy of arithmetic operations performed on them.

In legitimate latent trait measurement, summation is explicitly rejected as a basis for measurement. Models such as the Rasch model exist precisely because raw scores cannot be treated as measures. Rasch analysis tests unidimensionality, calibrates item difficulty and person ability on a common scale, and transforms ordinal responses into interval-level measures under strict conditions. Even then, the resulting scale is a logit-based ratio scale only if the model fits. Summation alone never suffices.

The claim that summations of subjective responses are ratio measures is therefore false in a categorical sense. Such summations produce convenient numbers, not quantities. Treating them as ratio measures enables illegitimate arithmetic and undermines any claim to scientific measurement.

Statement: *“Meeting the axioms of representational measurement is required for arithmetic.”*
Classification: TRUE

Arithmetic operations are meaningful only when the quantities involved satisfy the axioms of representational measurement theory. RMT specifies the conditions under which numerical representations preserve the structure of empirical attributes. These conditions determine which arithmetic operations are permissible. Without meeting the relevant axioms, arithmetic ceases to represent anything in the real world and becomes purely symbolic.

RMT establishes that different scale types support different operations. Nominal scales permit only classification. Ordinal scales permit ranking but not addition. Interval scales permit addition and subtraction but not multiplication or division. Ratio scales permit the full range of arithmetic operations. These distinctions are not conventions; they are logical consequences of how

empirical relations can be mapped onto numbers. Arithmetic does not exist independently of measurement. It is constrained by it.

The axioms required for a given scale type are therefore preconditions for arithmetic. Equal intervals are required for addition and subtraction to be meaningful. A true zero is required for multiplication, division, and ratio statements. Unidimensionality is required for any arithmetic to correspond to variation along a single attribute. Invariance is required so that arithmetic results do not depend on the particular items, respondents, or contexts used to generate the numbers. If these axioms are not satisfied, arithmetic results cannot be interpreted as statements about the attribute of interest.

In scientific practice, these constraints are enforced implicitly through dimensional analysis and unit consistency. Physicists do not add quantities measured in different dimensions, nor do they multiply variables without establishing their scale properties. These practices reflect an underlying commitment to RMT axioms, even when the theory is not named explicitly.

In health technology assessment, this ordering is routinely reversed. Arithmetic operations are specified first, and scale properties are assumed as needed to justify them. Utilities are treated as interval or ratio measures because cost-effectiveness analysis requires subtraction, multiplication, and division. This inversion violates RMT axioms. No amount of statistical modeling can compensate for arithmetic performed on quantities that lack the required measurement properties.

The requirement that RMT axioms be met before arithmetic is therefore foundational, not optional. It marks the boundary between scientific quantification and numerical storytelling. Once arithmetic is detached from measurement axioms, numbers no longer represent attributes; they merely decorate decisions. Recognizing this requirement exposes many familiar calculations in HTA as not just questionable, but illegitimate.

Statement: *"There are only two classes of measurement: linear ratio and Rasch logit ratio."*
Classification: TRUE

When measurement is understood in its strict scientific sense, as the lawful mapping of empirical attributes onto numbers such that arithmetic preserves meaning, there are only two defensible classes of measurement: linear ratio measurement for manifest attributes, and Rasch logit ratio measurement for latent attributes. This claim does not deny the existence of nominal, ordinal, or interval scales as classificatory or descriptive devices. It asserts that only ratio-scale structures support genuine measurement capable of underpinning arithmetic claims.

Linear ratio measurement applies to manifest attributes that are directly observable and concatenable, such as time, length, mass, counts, or resource use. These attributes possess a natural ordering, equal intervals, and a true zero representing absence. Because these properties are grounded in the empirical structure of the attribute itself, arithmetic operations, addition, subtraction, multiplication, and division, are meaningful. Linear ratio measures support

statements about magnitude, proportion, and change without ambiguity. They form the backbone of the physical sciences and of any empirical discipline that makes quantitative claims about observable phenomena.

Latent attributes are fundamentally different. Traits such as pain severity, functional ability, need fulfillment, or quality of life are not directly observable or concatenable. They cannot be measured by simple counting or direct comparison. For such attributes, classical summation of responses fails because ordinal observations do not possess equal intervals or a true zero. Rasch measurement provides the only scientifically coherent solution to this problem. It specifies the conditions under which ordinal responses can be transformed into measures by constructing a probabilistic measurement model that jointly estimates item difficulty and person ability on a common scale.

Crucially, Rasch measurement yields a logit scale that has ratio properties. The logit expresses the natural logarithm of the odds of a successful response, anchored in probabilistic structure rather than arbitrary scoring. Differences on the Rasch scale represent constant relative differences in the underlying latent trait, and the zero point is defined in relation to the probabilistic balance between person ability and item difficulty. This produces a ratio-scale measure, albeit one expressed logarithmically rather than linearly.

Interval scales, composite indices, weighted sums, and preference scores do not constitute measurement in this strict sense. They may produce ordered or numerically spaced values, but they lack either a true zero, invariance, unidimensionality, or lawful concatenation. As a result, arithmetic performed on them does not preserve empirical meaning. They describe, but they do not measure.

The claim that there are only two classes of measurement is therefore not reductive; it is clarifying. It draws a sharp boundary between quantities that can support scientific inference and those that cannot. In health technology assessment, recognizing this distinction is decisive. Manifest outcomes require linear ratio measures. Latent outcomes require Rasch logit ratio measures. Anything else is numerical representation without measurement.

Statement: *"Transforming subjective responses to interval measurement is only possible with Rasch rules"*

Classification: TRUE

Subjective responses are, by their nature, ordinal observations. They indicate order, more or less, better or worse, but they do not specify magnitude. Transforming such responses into measures requires a formal measurement model that can convert ordinal information into a scale with interval or ratio properties. Rasch measurement provides the only scientifically coherent set of rules for doing so. Without Rasch rules, subjective responses cannot be transformed into measures at all.

The core problem is that subjective responses do not possess equal intervals. A Likert response of "4" is not twice as much as a response of "2," nor is the difference between "2" and "3" necessarily equal to the difference between "4" and "5." Ordinal categories merely rank responses. Summing or averaging them does not create interval structure; it simply aggregates orderings. Any arithmetic performed on such aggregates assumes properties that have not been demonstrated.

Rasch measurement addresses this problem by specifying a probabilistic relationship between a person's level on a latent trait and the difficulty or intensity of an item. The Rasch model does not assume that responses are measures. It tests whether the pattern of responses can be explained by variation along a single latent dimension. Only when this condition is met can a scale be constructed. Unidimensionality, local independence, and invariance are not optional assumptions; they are requirements that must be satisfied by the data.

When data fit the Rasch model, ordinal responses can be transformed into a logit scale. The logit is the natural logarithm of the odds of a successful response. This transformation yields a scale with interval and ratio properties in probabilistic terms. Differences on the logit scale represent constant relative differences in the underlying latent trait, and comparisons are invariant across items and persons within the calibrated frame. No other approach to subjective measurement achieves this.

Alternative methods such as summated scores, factor scores, item response models that relax Rasch constraints, or regression-based scaling do not solve the measurement problem. They may improve prediction or fit, but they do not establish the conditions required for measurement. Relaxing Rasch requirements sacrifices invariance, which is the defining feature that distinguishes measurement from description.

In health technology assessment, subjective outcomes are ubiquitous, particularly in patient-reported outcomes and quality-of-life instruments. Treating raw or summed subjective responses as measures enables arithmetic that has no empirical meaning. Transforming subjective responses therefore requires Rasch rules. Without them, numbers may be produced, but measurement has not occurred.

Statement: *Summation of Likert question scores creates a ratio measure* **Classification: FALSE**

The claim that summing Likert question scores creates a ratio measure is false as a matter of elementary mathematics and representational measurement theory. Likert items generate **ordinal** data. They record ordered categories—such as "strongly disagree" to "strongly agree"—but they do not establish equal distances between categories, a true zero, or invariance across respondents. Summation does not repair these deficiencies. Adding ordinal numbers does not transform them into a quantitative measure any more than adding ranks in a race produces a measure of speed.

A ratio measure requires four conditions: unidimensionality, equal units, a meaningful zero, and invariance under admissible transformations. Likert responses satisfy none of these conditions. The numeric labels assigned to response categories are arbitrary; replacing 1–5 with 0–4 or 10–

50 preserves order but changes sums. If the numerical values can be altered without changing the empirical meaning of the responses, the resulting totals cannot represent quantities. Ratio measures are invariant under multiplication by a positive constant; Likert sums are not invariant under even simple relabeling.

Summation also assumes equal spacing between response categories. There is no empirical justification for assuming that the subjective distance between "agree" and "strongly agree" is the same as between "neutral" and "agree," either within or across respondents. Without equal units, addition is undefined. Treating these category labels as if they were distances substitutes convenience for measurement.

The absence of a true zero is decisive. Ratio scales require a zero point that represents the absence of the attribute. Likert scales have no such anchor. A total score of zero or the lowest possible sum does not represent "no satisfaction," "no pain," or "no quality of life." It merely represents selection of the lowest available category across items. Without a true zero, multiplication, division, and meaningful ratios are impossible.

Summation further compounds multidimensionality. Likert instruments almost always mix multiple attributes—frequency, intensity, emotional valence, functional impact—within and across items. Adding scores assumes these attributes are commensurable and contribute equally to a single latent variable. That assumption is rarely tested and almost never justified. Without demonstrated unidimensionality, the sum has no coherent empirical interpretation.

Only a lawful transformation—such as Rasch measurement—can convert ordinal responses into an interval-level scale by estimating item difficulty and person location under conditions of invariance. Mere summation performs no such transformation. It produces a larger ordinal number, not a measure. To treat Likert sums as ratio quantities is therefore not approximation; it is mathematical error.

Statement: *"The QALY is a dimensionally homogeneous measure."* **Classification: FALSE**

The QALY is not dimensionally homogeneous. Dimensional homogeneity requires that a quantity represent variation along a single empirical dimension, such that all components share the same dimensional meaning and can be combined without ambiguity. The QALY violates this requirement by construction. It combines time, a manifest ratio-scale quantity, with a preference-weighted index of health states that lacks a single dimension and lacks lawful measurement properties. The resulting product does not represent one attribute; it is a composite of incommensurate elements.

Dimensional homogeneity is a foundational constraint in all sciences that employ arithmetic. In physics and engineering, quantities can be added, multiplied, or divided only when their dimensions are compatible, and derived quantities have clearly defined dimensions. Velocity, for example, has the dimension of length per unit time; force has the dimension of mass times acceleration. These derived quantities are meaningful because their components are themselves

measures with known scale properties and because the combination preserves a coherent dimensional interpretation.

The QALY has no such coherence. Time is a ratio measure with a true zero and equal intervals. The utility component, however, is not a measure of a single attribute. It is derived from preference elicitation exercises applied to multidimensional health-state descriptions. These descriptions explicitly combine disparate domains such as mobility, pain, mood, and self-care. Weighting and summing these domains produces a scalar index, but not a dimension. A scalar is not the same as a dimensionally homogeneous quantity.

Multiplying time by this index does not create homogeneity. Time remains a duration; the utility score remains a preference-based summary. The product has no interpretable dimension. It is not "time," not "health," and not any recognized composite with lawful properties. The label "quality-adjusted life years" suggests a dimensional structure analogous to physical quantities, but this is rhetorical, not scientific. There is no empirical attribute called "quality-adjusted time" with a defined dimension that the QALY measures.

The problem is exacerbated by the lack of invariance in the utility component. Different instruments, valuation protocols, populations, and anchoring conventions produce different utility scales. If the same health state can have different numerical values depending on context, then the dimension being represented is not stable. Dimensional homogeneity requires invariance; without it, arithmetic combinations cannot preserve meaning.

The existence of negative QALYs in some valuation systems further exposes the incoherence. A dimensionally homogeneous ratio quantity cannot cross zero without representing absence of the attribute. Negative QALYs imply "negative health-time," a notion with no dimensional interpretation. This alone is sufficient to refute homogeneity.

The claim that the QALY is dimensionally homogeneous is therefore false. The QALY is a constructed index that combines heterogeneous elements under a single label. Treating it as a homogeneous quantity enables arithmetic that appears scientific but lacks any defensible dimensional meaning.

Statement: *"Claims for Cost-effectiveness fail the axioms of representational measurement*

Classification: TRUE

Cost-effectiveness claims, as routinely constructed in health technology assessment, fail the axioms of representational measurement theory. This failure is structural, not incidental. It arises because cost-effectiveness analysis performs arithmetic operations, most notably division, on quantities whose measurement properties do not support those operations. When RMT axioms are violated, the resulting numerical claims cannot represent empirical relationships and therefore cannot be true or false.

A cost-effectiveness claim typically takes the form of an incremental cost-effectiveness ratio (ICER), in which differences in costs are divided by differences in outcomes, most often QALYs. For such a ratio to be meaningful, both the numerator and denominator must be ratio-scale measures with clearly defined dimensional properties. Costs, when expressed as monetary expenditures over a defined period, can plausibly satisfy ratio-scale requirements. Outcomes in cost-effectiveness analysis do not.

As shown elsewhere, QALYs are not ratio measures. They lack a true zero, fail unidimensionality, admit negative values in some valuation systems, and are not dimensionally homogeneous. Because the denominator of the ICER is not a lawful measure, the ratio itself has no empirical meaning. Dividing a ratio-scale cost by a non-measure does not yield a meaningful rate or efficiency metric. It yields a number without representational content.

This failure cannot be repaired by appealing to convention, widespread use, or policy necessity. RMT axioms are not recommendations; they are logical conditions that determine when numbers can stand in for attributes. If those conditions are not met, arithmetic does not "approximately" work—it fails categorically. No amount of sensitivity analysis, probabilistic modeling, or threshold adjustment can rescue a ratio whose denominator does not measure anything.

Cost-effectiveness claims also fail invariance, another core RMT requirement. The numerical value of an ICER depends on the choice of instrument, valuation method, population preferences, modeling horizon, discount rate, and structural assumptions. A quantity whose value changes systematically with the method used to generate it cannot represent a stable empirical relation. It represents a modeling choice, not a property of the world.

Moreover, cost-effectiveness claims conflate multiple dimensions under a single numerical expression. They purport to represent "efficiency," but this efficiency is not an attribute that can be independently observed or measured. It is an artifact of combining heterogeneous components, costs, preferences, time, and modeling assumptions into a single ratio. RMT does not permit such composites to be treated as measures.

The conclusion follows directly from measurement theory: cost-effectiveness claims fail RMT axioms and therefore fail as scientific claims. They may function as administrative heuristics or decision rituals, but they do not measure efficiency, value, or benefit in any empirical sense. Once this is acknowledged, cost-effectiveness analysis must be reclassified, not as flawed science, but as non-measurement masquerading as quantification.

Statement: *"QALYs can be aggregated."*
Classification: FALSE

QALYs cannot be legitimately aggregated because aggregation presupposes that the quantities being summed are measures with the same dimensional meaning, scale properties, and invariance across persons and contexts. QALYs fail these requirements. As a result, summing QALYs across individuals, populations, or interventions does not produce a meaningful quantity. It produces a numerical total that lacks empirical interpretation.

Aggregation is a lawful operation only when the units being added represent the same attribute on the same scale. In the physical sciences, lengths can be summed because each unit of length represents the same dimension with equal intervals and a true zero. Counts can be aggregated because each count represents the presence of one identical unit. Aggregation preserves meaning because the underlying measurement structure is coherent. QALYs do not meet this standard.

The first failure is dimensional. A QALY is constructed by multiplying time by a preference-weighted index of health states. As established elsewhere, this index is not a measure of a single attribute and lacks dimensional homogeneity. Aggregating such products assumes that each QALY unit represents the same "amount of health" regardless of the health state, the individual, or the valuation method. This assumption has no empirical foundation. A QALY gained through pain relief is not dimensionally equivalent to a QALY gained through improved mobility or extended survival. They are numerically commensurated by convention, not by measurement.

Second, aggregation requires invariance. The meaning of a unit must be stable across persons and contexts. QALYs fail invariance because their values depend on preference elicitation methods, population samples, cultural context, and anchoring conventions. The same health state can generate different utility values across studies and jurisdictions. If the unit itself is unstable, aggregation is meaningless. One cannot sum quantities whose units shift with the method of estimation.

Third, aggregation presupposes a true zero. For a total to represent "more" of an attribute, zero must represent its absence. QALYs do not have a true zero. Zero QALYs reflects zero time, not zero health. Moreover, the existence of negative QALYs in some valuation systems undermines any claim that aggregation represents accumulation of an attribute. Aggregating quantities that can cross zero without representing absence destroys interpretability.

The ethical appeal of QALY aggregation—maximizing total health gain—does not rescue its measurement failure. Ethical arguments cannot substitute for measurement axioms. Summing non-measures does not become meaningful because the goal is morally attractive. It remains arithmetic without representational content.

The claim that QALYs can be aggregated is therefore false in a categorical sense. Aggregation assumes measurement. QALYs do not satisfy the conditions required to be measures. Summing them produces a number, but not a total of anything that exists in the empirical world.

Statement: *"Non-falsifiable claims should be rejected."*
Classification: TRUE

Non-falsifiable claims should be rejected because falsifiability is a necessary condition for any claim that purports to be scientific. This principle is not a matter of philosophical preference or methodological fashion; it is the criterion that distinguishes empirical claims from belief, speculation, or narrative. A claim that cannot, even in principle, be shown to be false cannot contribute to the growth of knowledge. It can be asserted, defended, or repeated, but it cannot be tested.

Falsifiability requires that a claim specify the conditions under which it would fail. This does not mean that the claim must be false, only that it must expose itself to the risk of refutation by observation. In the absence of such exposure, a claim is insulated from evidence. No amount of data can confirm or disconfirm it, because no conceivable outcome counts against it. Such claims are immune to correction and therefore incompatible with scientific inquiry.

In mature sciences, non-falsifiable claims are excluded by default. A physical theory that cannot generate testable predictions is not treated as provisional science; it is treated as metaphysics. An engineering model that cannot be validated against observable performance is rejected regardless of its elegance. This exclusion is not punitive. It is protective. It ensures that science remains a process of conjecture and refutation rather than accumulation of unfalsifiable assertions.

In health technology assessment, this boundary has been systematically eroded. Model-based lifetime cost-effectiveness claims are routinely presented as evidence even when they cannot be empirically tested. Claims extending decades into the future, dependent on unobservable counterfactuals, structural assumptions, and preference weights, are treated as if they were provisional truths. When challenged, they are defended not by proposing empirical tests, but by appealing to plausibility, expert consensus, or sensitivity analysis. None of these renders a claim falsifiable.

Sensitivity analysis, in particular, is often mistaken for falsification. Varying assumptions across ranges does not test whether a claim is true; it merely explores how outputs change when inputs are altered. A claim that survives sensitivity analysis is not confirmed; it is simply robust within its own speculative framework. If there is no possible observation that could refute the claim, robustness is irrelevant.

Accepting non-falsifiable claims has predictable consequences. It allows models to persist regardless of empirical failure, because failure cannot be defined. It shifts evaluation from evidence to craftsmanship, rewarding complexity rather than truth. Over time, it transforms HTA from an empirical discipline into a narrative one, where claims are judged by coherence and authority rather than testability.

The requirement to reject non-falsifiable claims is therefore not extreme. It is minimal. It sets the lowest bar for scientific legitimacy. Once this bar is removed, anything can be asserted as "evidence," and nothing can be conclusively rejected. In that environment, numbers cease to inform decisions and begin to legitimize them.

To insist that non-falsifiable claims be rejected is to insist that HTA choose science over storytelling.

Statement: *"Reference case simulations generate falsifiable claims."* **Classification: FALSE**

Reference case simulations do not generate falsifiable claims. They generate internally consistent numerical outputs whose validity depends entirely on assumptions that are not empirically testable as a whole. This distinction is critical. A claim is falsifiable only if there exists a

conceivable empirical observation that would contradict it. Reference case simulation outputs do not meet this criterion, because they are insulated from refutation by their construction.

A reference case simulation is not a hypothesis about the world; it is a conditional projection. It answers the question: *what would happen if a specified set of assumptions were true?* The assumptions typically include model structure, health-state transitions, extrapolated survival functions, preference weights, discount rates, and behavioral rules extending far beyond observed data. The resulting outputs of lifetime costs, QALYs, and ICERs are therefore logical consequences of assumptions, not empirical claims about observable outcomes.

Because these outputs depend on unobservable counterfactuals and long-term projections, there is no possible observation that could refute them. One cannot observe the lifetime trajectory of a counterfactual population that did not receive the intervention. One cannot empirically verify utilities projected decades into the future. When observed outcomes differ from model projections, the model is not falsified; it is revised. Assumptions are adjusted, horizons extended, or alternative scenarios introduced. This flexibility ensures survival of the model regardless of empirical discrepancy.

Sensitivity analysis does not restore falsifiability. Varying parameters across plausible ranges explores how outputs respond to assumption changes, but it does not define conditions under which the claim fails. A model that produces a favorable ICER across a wide range of assumptions is not confirmed; it is merely robust within its own speculative space. Robustness is not falsification. A claim that cannot fail cannot be tested.

True falsifiable claims in HTA would be time-bounded, population-specific, and linked to observable outcomes. For example, a claim that an intervention reduces hospital days by a specified amount over twelve months in a defined population can be empirically evaluated and refuted. Reference case simulations do not make such claims. They substitute projected lifetime constructs for testable propositions, thereby avoiding exposure to empirical risk.

The reference case framework further entrenches non-falsifiability by standardizing assumptions. Once a particular modeling architecture is mandated, deviations are interpreted as methodological noncompliance rather than as opportunities for empirical challenge. This transforms critique into process policing rather than hypothesis testing.

The claim that reference case simulations generate falsifiable claims is therefore false in a categorical sense. They generate narratives constrained by assumptions, not hypotheses exposed to refutation. Treating such outputs as evidence does not approximate science; it replaces it with disciplined speculation.

Statement: *“The logit is the natural logarithm of the odds ratio.”* **Classification: TRUE**

The logit is, by definition, the natural logarithm of the odds. This is a precise mathematical identity, not a modeling convention or a statistical approximation. If the probability of an event is denoted by p, the odds of that event are defined as $p / (1 - p)$. The logit is then defined as $ln[p$

/ (1 − p)], where *ln* denotes the natural logarithm. There is no alternative definition. Any quantity described as a logit must satisfy this relationship.

This definition has important implications for measurement. Odds express relative likelihood: how much more likely an event is to occur than not to occur. Taking the natural logarithm of the odds transforms a bounded probability scale, which lies between 0 and 1, into an unbounded continuous scale extending from negative infinity to positive infinity. This transformation is not cosmetic. It creates a scale on which equal differences correspond to constant multiplicative differences in the underlying odds. That property is what gives the logit its ratio-scale interpretation.

In Rasch measurement, the logit plays a central role because it links observable responses to an underlying latent trait through a probabilistic structure. The Rasch model expresses the probability of a successful response as a logistic function of the difference between person ability and item difficulty. When this probability is transformed into logits, the resulting scale represents the natural logarithm of the odds of success. Differences on the logit scale therefore represent constant relative differences in odds, regardless of where they occur on the scale.

This is why the logit scale has ratio properties in probabilistic terms. A difference of one logit corresponds to the same multiplicative change in odds anywhere on the scale. A two-logit difference represents the square of that change, and so on. These multiplicative relations are invariant, which is the defining characteristic of ratio measurement. Although the scale is logarithmic rather than linear, it is still a ratio scale because ratios of odds are preserved through addition and subtraction on the logit scale.

Confusion sometimes arises because the logit is discussed in statistical contexts without attention to its measurement implications. In logistic regression, for example, logits are often treated as mere coefficients. But their meaning remains the same: they are natural logarithms of odds ratios. This meaning does not disappear when the logit is embedded in a model.

In the context of health measurement, recognizing what a logit is matters because it distinguishes Rasch-based measurement from ad hoc scoring. When subjective responses are transformed into logits under Rasch rules, the resulting values are not arbitrary scores. They are measures grounded in a defined probabilistic structure. The statement that the logit is the natural logarithm of the odds ratio is therefore unambiguously true, and it underpins the claim that Rasch measurement yields lawful, ratio-scale quantities for latent traits.

Statement: *“The Rasch logit ratio scale is the only basis for assessing therapy impact for latent traits”*
Classification: TRUE

Latent trait impact can be meaningfully assessed only if the latent trait itself has been measured. This requirement immediately rules out most approaches used in health technology assessment and outcomes research, because latent traits such as pain severity, functional ability, fatigue, need fulfillment, or need fulfillment are not directly observable. They cannot be counted,

concatenated, or expressed on linear ratio scales. Any claim about change or impact on a latent trait therefore presupposes a valid measurement framework capable of transforming subjective observations into a lawful measure. The Rasch logit ratio scale provides the only scientifically coherent basis for doing so.

The defining challenge of latent traits is that observations are indirect. Responses to questionnaire items are ordinal indicators of an underlying construct, not measures of magnitude. Ordinal data do not support arithmetic, and differences between raw scores do not correspond to equal differences in the latent trait. Without transformation, it is impossible to say whether a change of two points reflects a small or large impact, or whether the same change has the same meaning across persons or contexts. Claims of impact under these conditions are descriptive at best and meaningless at worst.

Rasch measurement resolves this problem by imposing strict measurement requirements rather than assuming them. The Rasch model specifies a probabilistic relationship between a person's level on a latent trait and the difficulty of an item. When data fit the model, ordinal responses can be transformed into a logit scale representing the natural logarithm of the odds of a successful response. This transformation yields a scale with ratio properties in probabilistic terms: equal differences on the scale correspond to constant relative differences in the underlying trait.

Crucially, Rasch measurement enforces invariance. Person measures are independent of the particular items used, and item calibrations are independent of the particular sample, within the limits of model fit. This invariance is what makes impact claims meaningful. Without it, observed change could reflect changes in item functioning, respondent interpretation, or sample composition rather than true change in the latent trait. Rasch measurement uniquely separates these sources of variation.

Alternative approaches such as summated scores, factor scores, or item response models that relax Rasch constraints do not establish measurement. They may improve statistical fit or predictive accuracy, but they sacrifice invariance. Once invariance is lost, the resulting numbers cannot support claims of impact, because changes cannot be attributed unambiguously to the latent trait itself.

In health technology assessment, latent trait impact claims are ubiquitous, particularly in patient-reported outcomes. Treating raw or summed scores as measures enables arithmetic comparisons that have no empirical foundation. Only the Rasch logit ratio scale provides a defensible basis for quantifying latent traits and evaluating impact. Without it, claims of improvement, deterioration, or comparative benefit remain narratives, not measurements.

Statement: *"A linear ratio scale for manifest claims can always be combined with a logit scale."*
Classification: FALSE

Linear ratio scales cannot always be combined with logit scales, and treating such combinations as automatically legitimate reflects a misunderstanding of both measurement theory and

dimensional coherence. The fact that both linear ratio scales and Rasch logit scales possess ratio properties does not mean that they can be freely multiplied, added, or otherwise combined. Ratio status is a necessary condition for arithmetic, but it is not sufficient. Dimensional compatibility is also required.

A linear ratio scale represents a manifest attribute with a natural unit and a true zero grounded in the empirical structure of the attribute itself. Time, length, counts, and resource use fall into this category. Arithmetic on these quantities is meaningful because concatenation and proportional comparison correspond directly to real-world relations. The units are interpretable independently of any probabilistic model.

A Rasch logit scale, by contrast, is a probabilistic ratio scale. It represents a latent trait in terms of the natural logarithm of odds. Differences on the logit scale correspond to constant relative differences in odds, not constant additive differences in a manifest attribute. The zero point on a Rasch scale is defined relationally—where person ability equals item difficulty—not as an absence of the trait. Although the logit scale supports ratio interpretation in probabilistic terms, its units are fundamentally different from those of linear ratio scales.

Because of this difference, combining linear ratio and logit scales is not automatically meaningful. Multiplication or addition across scales presupposes that the resulting quantity represents a coherent empirical attribute. In most cases, it does not. For example, multiplying time (a linear ratio measure) by a Rasch logit score does not yield a quantity with a clear dimensional interpretation. The product is neither time nor latent trait, nor any recognized composite grounded in empirical structure. The fact that both inputs have ratio properties does not rescue the operation.

This point is often misunderstood because HTA practice assumes that ratio scales are universally compatible. In reality, dimensional analysis still applies. Even in physics, ratio-scale quantities cannot be arbitrarily combined. Time can be multiplied by velocity to produce distance, but only because the dimensional relationship is defined and empirically meaningful. Multiplying time by electric charge does not produce a meaningful quantity unless a theory specifies what that product represents. Measurement theory does not license free combination; it licenses only combinations that preserve dimensional meaning.

In health technology assessment, the false belief that linear ratio and logit scales can always be combined underlies attempts to merge manifest outcomes with latent trait measures into single summary quantities. This repeats the error of the QALY in a different guise. The presence of ratio properties is treated as sufficient justification for arithmetic, while dimensional coherence is ignored.

The statement is therefore false. Linear ratio scales and Rasch logit ratio scales can be combined only under explicitly defined theoretical conditions that specify the meaning of the resulting quantity. In the absence of such conditions, combination produces numbers without interpretability. Measurement permits arithmetic, but it does not abolish dimensional discipline.

Statement: *"The outcome of interest for latent traits is the possession of that trait."*
Classification: TRUE

For latent traits, the outcome of interest is possession of the trait itself, not a score, index, or summary statistic derived from responses. This follows directly from what a latent trait is. A latent trait is an unobservable attribute inferred from patterns of observable behavior or responses, such as pain severity, functional ability, fatigue, or need fulfillment. Because the trait is not directly observable, the only scientifically meaningful outcome is whether and to what extent the trait is possessed by an individual.

Measurement, in this context, is not about tallying responses or optimizing prediction. It is about establishing a scale on which possession of the latent trait can be located and compared. The object of inference is the person's position on the latent continuum. Any numerical value assigned has meaning only insofar as it represents that position. Scores that do not correspond to trait possession are not outcomes; they are artifacts of scoring rules.

This distinction is crucial. Raw questionnaire totals, factor scores, or preference indices are often treated as outcomes in their own right. But such quantities do not represent possession unless they are derived from a measurement model that enforces unidimensionality, invariance, and lawful transformation. Without these properties, changes in scores may reflect item composition, response styles, or contextual effects rather than changes in the underlying trait. In those cases, the "outcome" is ambiguous, and impact claims are uninterpretable.

Rasch measurement makes this explicit. The Rasch model estimates person ability (or trait level) and item difficulty on the same scale, allowing statements about possession to be made invariantly across items and samples. A person's measure represents their degree of possession of the latent trait relative to a calibrated frame of reference. Change over time, group differences, and treatment effects are meaningful only because they are interpreted as changes in possession, not as changes in raw scores.

Importantly, possession does not imply a binary yes/no state. Latent traits are typically continuous, and possession varies in degree. The outcome of interest is therefore the magnitude of possession on a valid measurement scale. What matters is whether an intervention increases, decreases, or otherwise alters possession of the trait, and by how much, in a way that is invariant and interpretable.

In health technology assessment, failure to recognize this leads to category errors. Outcomes are reported as changes in questionnaire totals or preference scores, and these are treated as if they directly represented benefit. They do not. Unless the numbers correspond to possession of a latent trait, the outcome has not been measured.

The statement is therefore true. For latent constructs, the only meaningful outcome is possession of the trait itself. Everything else is a proxy—and without lawful measurement, a misleading one.

Statement: *"The Rasch rules for measurement are identical to the axioms of representational measurement."*
Classification: TRUE

Rasch rules are the operational realization of representational measurement theory axioms for latent traits. They do not supplement RMT, relax it, or approximate it; they implement it. Where RMT specifies the conditions under which numerical representations can meaningfully stand in for empirical attributes, Rasch measurement provides the only known model that enforces those conditions for unobservable constructs.

RMT establishes that measurement requires more than numerical assignment. It requires unidimensionality, invariance, and a scale structure that supports lawful transformations. For latent traits, these requirements are especially demanding because the attribute cannot be observed or concatenated directly. Any claim to measurement must therefore demonstrate that observed responses can be explained by variation along a single latent dimension and that this explanation is invariant across persons and items. Rasch rules do exactly this.

Unidimensionality is enforced by the Rasch requirement that a single latent variable explains response probabilities. If responses cannot be modeled as a function of one trait, measurement fails. This is not a diagnostic add-on; it is a condition of the model. Data that violate unidimensionality are rejected as non-measuring. This corresponds directly to RMT's requirement that a measure represent only one attribute.

Invariance is the core of Rasch measurement and the defining feature of measurement in RMT. Rasch rules require that person measures be independent of the particular items used and that item calibrations be independent of the particular sample, within a defined frame of reference. This mutual invariance is precisely what RMT demands when it specifies that measurement must preserve empirical relations under admissible transformations. Without invariance, numbers may predict, but they do not measure.

Scale structure is also enforced. Rasch measurement yields a logit scale, the natural logarithm of the odds, which has ratio properties in probabilistic terms. Equal differences represent constant relative differences in the latent trait, satisfying the RMT requirement that numerical differences correspond to meaningful differences in the attribute. This is not assumed; it is derived from the model's probabilistic structure.

Crucially, Rasch rules are falsifiable. If data do not fit the model, measurement has not occurred. This aligns with RMT's insistence that measurement claims be testable and refutable. Other approaches to latent variables often relax constraints to improve statistical fit, but doing so abandons measurement in favor of description or prediction. Rasch measurement refuses that trade-off.

The equivalence between Rasch rules and RMT axioms explains why no alternative model has succeeded in establishing measurement for latent traits. Where Rasch rules are violated, RMT axioms are violated. Where Rasch rules hold, RMT requirements are satisfied.

The statement is therefore true. Rasch rules are not one approach among many; they are the applied form of representational measurement theory for latent constructs. Without them, claims to measurement are empty.

PRINT