CellPress
OPEN ACCESS

Opinion

Measurement theory and paleobiology

Kjetil Lysne Voje [ID], [1,*] James G. Saulsbury, [1] Jostein Starrfelt, [1] Daniel Varajão Latorre, [2] Alexis Rojas, [3] Vilde Bruhn Kinneberg, [1] Lee Hsiang Liow, [1] Connor J. Wilson, [1,4] Erin E. Saupe, [5] and Mark Grabowski[6]

Measurement theory, a branch of applied mathematics, offers guiding principles for extracting meaning from empirical observations and is applicable to any science involving measurements. Measurement theory is highly relevant in paleobiology because statistical approaches assuming ratio-scaled variables are commonly used on data belonging to nominal and ordinal scale types. We provide an informal introduction to representational measurement theory and argue for its importance in robust scientific inquiry. Although measurement theory is widely applicable in paleobiology research, we use the study of disparity to illustrate measurement theoretical challenges in the quantitative study of the fossil record. Respecting the inherent properties of different measurements enables meaningful inferences about evolutionary and ecological processes from paleontological data.

Measurements as a means to learn about the empirical world

Measurements represent a particular quality or aspect of the entity being measured, allowing us to make inferences about the empirical world based on our data. For example, instead of eyeballing a sample of ammonites to assess which is the largest, we can compare measurements of their diameter. How to extract meaning from measurements is the domain of **measurement theory** (see Glossary) [1,2].

Measurement theoretical principles are applicable to any quantitative science and are by no means constrained to the study of evolutionary processes by analyzing the fossil record. Why, then, the focus on measurement theory in paleobiology? Measurement theory has generally received little attention in the biological sciences, with a few notable exceptions [3–7]. By extension, discussions on the application of measurement theory to paleobiology are rare (but see [8,9]). Measurement theory is, in our view, highly relevant to consider in paleobiology because key research questions in the field often involve applying the same statistical models to measurements that belong to different scale types (Box 1). This practice makes paleobiology prone to the violation of several key measurement theoretical principles, which may impact inferences made from the fossil record. These violations can occur across paleobiological subdisciplines, from paleobiogeography to phylogenetics to morphometrics, among others.

The field of paleobiology emerged during the 1950s and 1960s, when some paleontologists began to shift their focus from a more descriptive science of life in the rock record to a more quantitative and model-based effort to use fossil data to infer evolutionary patterns and processes [10–12]. Paleobiology has been a huge success, bringing important insights to further the development of evolutionary biology [13]. Work within paleobiology has seen a steady rise in the application of more sophisticated statistical machinery on data collected from the fossil record. Students of evolutionary and ecological patterns and processes in the fossil record therefore have an increasingly sophisticated toolbox to choose from in their work. The application of statistical tools to fossil data

[1]Natural History Museum, University of Oslo, Oslo, Norway
[2]Department of Natural Sciences, Manchester Metropolitan University, Manchester, UK
[3]Department of Computer Science, University of Helsinki, Helsinki, Finland
[4]School of Geography and the Environment, University of Oxford, Oxford, UK
[5]Department of Earth Sciences, University of Oxford, Oxford, UK
[6]Research Centre in Evolutionary Anthropology and Paleoecology, Liverpool John Moores University, Liverpool, UK

*Correspondence:
k.l.voje@nhm.uio.no (K.L. Voje).

creates exciting possibilities for new insight, but also some potential measurement theoretical pitfalls.

Here, we begin by introducing some fundamental principles in measurement theory and explore their relevance to paleobiological studies. We show that although measurement theory is not typically applied explicitly when planning or conducting research, we are consciously or unconsciously making measurement theoretical decisions that influence the accuracy, interpretability, and **meaningfulness** of our work. Therefore, measurement theory can provide a framework to aid our thinking, help us to make our assumptions explicit, and enable us to conduct better and more meaningful science.

An informal introduction to measurement theory

We provide a brief and informal introduction to representational measurement theory in this section. For more comprehensive, formal, and axiomatic treatments, we recommend [2,14,15].

Representational measurement theory operates under the core assumption that a measurement represents a specific aspect of the entity that is being measured, such as the length in millimeters (aspect) of a trilobite (entity of interest). It is therefore worthwhile to make a distinction between values and measurements. Values are not automatically measurements but can become measurements with additional information regarding what aspect of reality they describe. Assigning values to aspects of an entity according to rules is what defines a measurement [1]. The same value can therefore hold different types of information, depending on which aspect of the entity they represent, and will differ in the permitted meaningful comparisons that can be made (Box 1).

For example, we might represent the presence or absence of spikes on the pygidium of a trilobite (Phylum Arthropoda) using the values 0 and 1. We can calculate the frequency of specimens with spikes on the pygidium among our trilobite samples by the frequency of scores of 1 for this measurement. We could furthermore assign values of 0, 1, or 2 to differentiate the lateral furrow development as absent, partially developed, and present for trilobite specimens. The number of specimens in each category allows us to draw meaningful conclusions about the most common category in our sample. To calculate the average size of trilobites in our sample, we could measure the lengths of our specimens with a ruler and calculate the arithmetic mean. These measurements describe different aspects of the trilobite, and we assign values according to rules for each aspect of the trilobite we wish to study (e.g., presence/absence of spikes, the size of specimens). Inferences based on measurements enable us to make meaningful statements as long as there is a clear link between our measurements and the reality that they represent.

Respecting scale types

The numbers describing the three traits in our trilobite sample (presence/absence of spikes, extent of lateral furrow development, and lengths of specimens) have different properties. In the context of measurement theory, we say that these measurements belong to different scale types [1,2] (Box 1). The definition of a scale type is tightly linked to which comparisons of the values are meaningful and how values can be changed or transformed without disrupting the connection between the values and the aspect of reality they embody.

Measurements that essentially function as labels belong to a nominal scale type. For example, measurements of the presence or absence of a trait (e.g., a spike on the pygidium of a trilobite) can be represented by any two values or symbols, so long as each state of the trait is represented by a unique 'tag.' Any one-to-one mapping – that is, substituting a set of symbols/values with another set of symbols/values – will preserve the link between the measurements and the reality

Glossary

Allowed transformation: transformations that preserve the link between the measurements and the aspect of reality they represent; also called 'permissible transformations.' A disconnecting transformation breaks the link between the measurement and the aspect of reality it represents. Disconnecting transformations are a cause of meaningless statements.

Meaningfulness: an inference/ statement based on measurements is meaningful if it is invariant to allowed transformations, given the scale type of the measurements. Conclusions based on numerical artifacts are the opposite of meaningful statements.

Representational measurement theory: theory about the relationship between values and the aspect of reality they represent. A key principle in measurement theory is to ensure that any inference we make based on measurements should also meaningfully apply to the entities the measurements represent.

Scale type: categories to which different measures belong. Every measurement belongs to a scale type. Several different scale types exist. Measurements belonging to the same scale type share how they can be changed (transformed) without losing their connection to the aspect of reality they represent (see Box 1). Measurements belonging to the same scale type also share the same set of meaningful comparisons that can be made among the measurements. Not all empirical/mathematical operations can be meaningfully conducted on measurements belonging to different scale types. The scale type thus constrains the types of statistics that can be meaningfully applied to the measurements.

Theoretical context: the ultimate reason why we conduct specific measurements. Measurements are collected to investigate a specific hypothesis within a given theoretical context. The theoretical context dictates the scale type our measurements should belong to.

they represent. The only meaningful comparison of two numbers/symbols on this scale type is whether they are the same. Two trilobites are either identical or distinct when it comes to the presence or absence of the spike.

Measurements that are ordered labels belong to an ordinal scale type. Ordinal variables convey order but not magnitude between values. Thus, a larger value on an ordinal scale means larger but does not convey information about how much larger. Any transformation that preserves order can be applied to values without breaking their link to reality. In other words, using the values 0, 1, and 2 or 0, 10, and 100; the letters a, b, and c; or the labels small, medium, and large are all equally valid. On an ordinal scale, we can assess whether a trilobite specimen should be classified as larger or smaller (or equal) compared with a different specimen, but we cannot say how much larger or smaller.

Box 1. Scale types

Table I. Common scale types in paleobiology

Meaningful comparisons[a]	Scale type	How to identify[b]?	Examples in paleobiology	Allowed transformations[c]	Example transformation	Common meaningful statistics[d]	Meaningless statistics
Equivalence	Nominal	Labels	Species names, presence/absence of character	Any one-to-one mapping	{Presence, absence} = {0, 1}	Count and frequency of cases, mode	Distances
Equivalence, order	Ordinal	Ordered labels	Preservation categories, size categories, extent of trait development	Substitutions that maintain order	{Poor, medium, excellent} = {0,1,2}	Median, percentiles	Distances, arithmetic and geometric means, variance
Equivalence, order, differences	Interval	Equally spaced ordered units	Relative size, dates	Any linear transformation	Fractionation of oxygen isotopes between calcium carbonate and water to temperature in Celsius [70].	Mean, standard deviation	Geometric mean
Equivalence, order, differences, ratios	Ratio	Equally spaced ordered units with a meaningful zero	Length, width	Multiplication by a constant	100 g = 0.1 kg = 0.22 lb	Log transformation, geometric and arithmetic means	

For each scale type, columns include types of comparisons/basic empirical operations that can be performed on measurements of that scale type (meaningful comparisons), the type of information each scale type contains (how to identify?), common examples in paleobiology, allowed or permitted transformations and examples of these transformations, meaningful statistics commonly performed in paleobiology, and statistics that do not make sense to apply to measurements of a given scale type.
[a]Meaningful comparisons/basic empirical operations intended to be performed with the measurements define the scale type.
[b]See also Box 3.
[c]Transformations that preserve the link between the empirical structure represented by the measurements.
[d]Incomplete list of statistics that return the same answer, given the allowed transformations.

Scale types (Table I) form a hierarchy based on the amount of information each type contains (see also Figure I). Stronger scale types contain more information than weaker scale types. For example, measurements on the nominal scale only contain information about identity (labels). Measurements on an ordinal scale include both labels and the order of the labels – these measurements have more information and are thus on a stronger scale type than nominal measures. As discussed in the main text, moving from stronger to weaker scale types is often possible, but the opposite conversion is not. For example, ranking ratio scale traits will convert them to the ordinal scale, but information cannot be added to measurements of a weaker scale type to convert them to a stronger scale. Mixing scale types and analyzing them together under the assumption that they contain the same type of information can have unforeseen consequences and may lead to meaningless results.

Statistics themselves are just 'machines' for producing results – they have no understanding of the meaning of a measurement such as '5' and whether this measurement is a category labeled '5' (nominal scale), whether it denotes a larger measurement than measurements labeled '4' and a smaller measurement than those labeled '6' (ordinal scale), or whether it means the measurement is one unit larger and one unit smaller than 4 and 6, respectively (ratio scale). Thus, the scale type to which the data belongs can have monumental consequences for the type of statistical tests appropriate to perform and whether the results are meaningful.
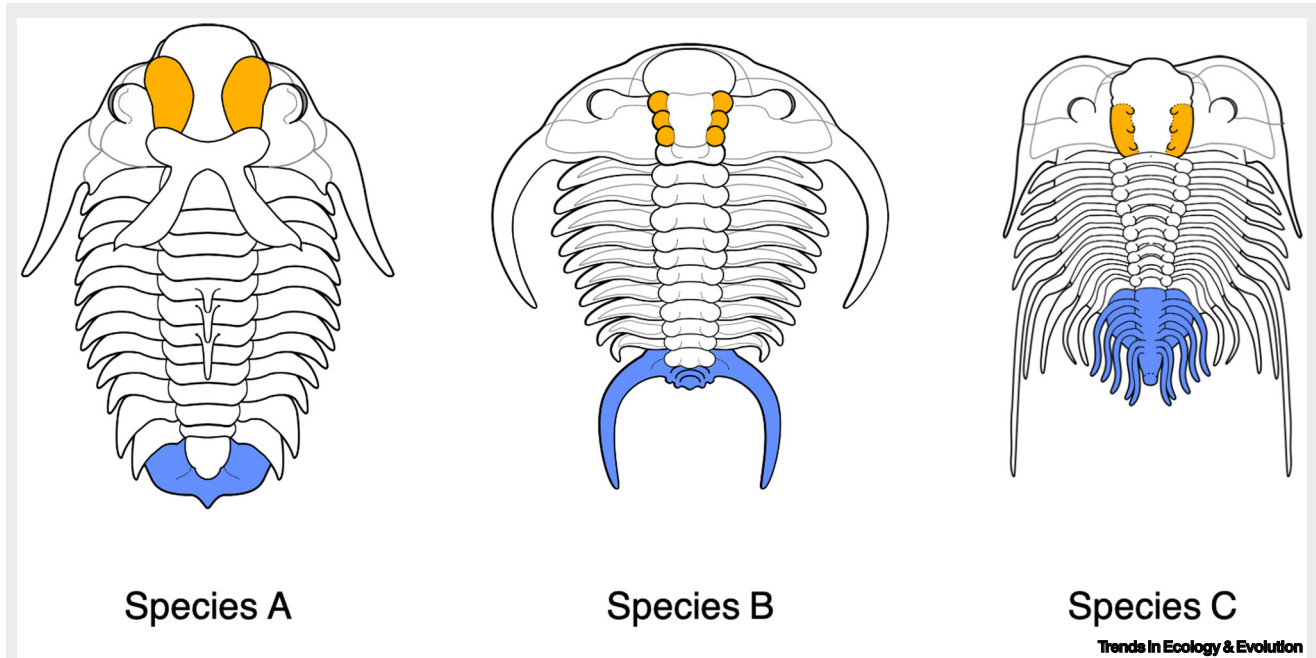
Species A   Species B   Species C

Trends in Ecology & Evolution

Figure I. Relationship between empirical comparisons and measurements showing the different scale types of each measurement. Note that nominal and ordinal type variables can be transformed into numeric symbols – e.g., 0/1 for present/absent or 0/1/2 for absent/partially developed/present. However, taking the mean of measurements on an ordinal variable (e.g., absent, absent, present, present, partially developed) for lateral glabellar furrows measured on a sample of five individuals is meaningless, even if the symbols are reclassified as numbers (e.g., 0, 0, 2, 2, 1).

Measurements where the difference between numbers is meaningful and where the number 0 represents a complete absence of the entity being measured belong to the ratio scale type. The length of a trilobite measured in millimeters is an example of a ratio scale variable. Multiplying a series of numbers on a ratio scale by a constant maintains the ratio between the measurements; for example, a comparison in millimeters and centimeters yields exactly the same results. When comparing two trilobites, 'twice as large' and '10% smaller' are accordingly meaningful statements when measurements belong to a ratio scale.

Measurements can belong to scale types other than nominal, ordinal, and ratio scales, but these three scale types are frequently collected and analyzed in paleobiological studies. (Box 1 explains scale types in more detail.) The fact that measurements belong to different scale types may seem trivial, but this insight is critical because the scale type constrains how measurements can be transformed and analyzed without breaking the link to the empirical entities they represent.

Meaningfulness

Meaningful statements rely on the premise that the outcome of an analysis should not depend on the values used to represent the traits being studied (Figure 1). If, for example, a nominal trait is represented with 0 and 1, 10 and 100, or A and B, the statements should be the same, regardless of which encoding is used. Similarly, an inference based on measurements is meaningful if it remains unchanged after transformations that preserve the link between the measurement and the reality they represent. For example, the only meaningful empirical operation that can be performed on measures of a nominal scale type is assessing equivalence. Labeling the presence or absence of a spike on the pygidium of a trilobite using the numbers 1 and 0 enables us to
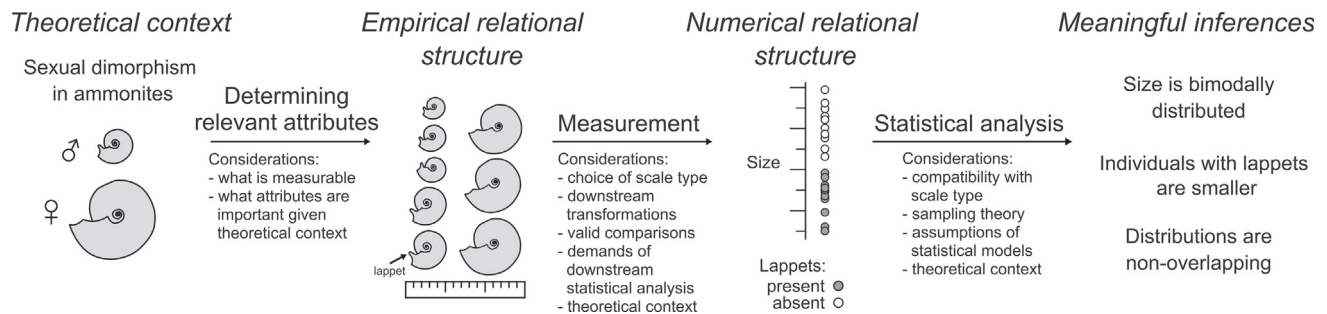
Figure 1. The role of measurement in quantitative inference. Illustration of measurement theoretical principles using sexual dimorphism in fossil ammonites, following [69] in hypothesizing that small, lappet-bearing individuals are males. Lappets are extensions on either side of the shell aperture in ammonites. Reaching the (meaningful) conclusions on the right first requires determining the measurable attributes of empirical objects that are relevant to the research questions, given the theoretical context. Selected attributes of those objects (shell diameter and presence/absence of lappets) are then assigned values according to rules – i.e., measured. These measurements are now representations of the relevant attributes. Here, size is a ratio scale measurement, whereas the presence or absence of lappets is on the nominal scale. The choice of measurements determines what downstream transformations can be made on those measurements and what valid comparisons can be made between them. Finally, measurements are subjected to statistical operations whose validity again depends on the measurement scale type. Most emphasis is usually placed on this final step – drawing conclusions from statistical analyses of measurements – but a measurement theoretical mistake in any previous step can lead to meaningless inference. Note that every step should be guided by the theoretical context.

compare whether particular specimens have a spike. Any other comparison based on these measurements (e.g., order or differences) is not meaningful, because this demands the measurements possess information they do not contain.

One can assess if the size of trilobite specimens scored using an ordinal measure is equivalent, larger, or smaller. However, it is meaningless to compute the mean and variance of such ordinal measurements because the magnitudes between numbers of ordinal measures have no connection to reality. For example, consider two samples of trilobites, one with 11, 13, and 6 specimens of small, medium, and large specimens, respectively, and another with 14, 8, and 8 specimens in the small, medium, and large size categories, respectively. If we use an ordinal scale of 0, 1, and 2 to score the categories, we obtain means of 0.83 and 0.80 for the first and second samples, which could easily (but erroneously) be interpreted as indicating larger specimens in the first compared with the second sample. However, because any numerical system that preserves order can be applied to an ordinal measurement, we could have used the equally arbitrary numbers 0, 10, and 100 to score the three ordered states, in which case, we would have obtained means of 24.3 and 29.3 for the first and second samples, changing the samples with the largest average. Instead, using the median will show that the first sample has a central tendency for larger specimens, regardless of the arbitrary numbers used to designate the size of each category. Medians or modes for central tendencies and ranges of values for variability are meaningful statistics on ordinal traits, but means and variances are not (Box 1).

Conclusions that are based on numerical artifacts, as illustrated above, are not meaningful. Respecting the scale type of our measurements in downstream analyses will ensure meaningful inference. Failure to do so may have the opposite effect. However, note that meaningfulness differs from truth. For example, the statement 'an adult mouse is heavier in grams than an adult elephant measured in grams' is untrue but meaningful because the mass of the two animals can be compared using weight measured in grams. It is also important to distinguish between meaningfulness and biological insight. On a ratio scale, equal deviations represent equivalent differences, which means that species of elephants are expected to evolve faster in size than species of mice in a given time interval, because 1 g gained by both species is treated as equivalent. It is meaningful from a measurement theoretical perspective to estimate and compare

rates of phenotypic change on a ratio scale [16], but it can be argued that comparing rates of change on log-transformed ratio scale variables (where equal deviations represent differences in equal proportions) is more sensible and interesting, given the multiplicative nature of animal growth (see also [17]).

The theoretical context of a study

Collecting measurements can be a tedious and expensive task, and researchers therefore typically have a specific goal in mind when gathering data. The purpose of a study, including the theoretical tradition upon which the work builds, is known as the **theoretical context** [4]. The theoretical context determines which measurement scales we can use to answer the research questions at hand. For instance, if the theoretical context is sexual selection and if the goal of a specific study is to assess whether the average size of male and female ammonites differ, ratio scale measurements are required because they provide interpretable magnitudes between numbers. Ordinal scale measurements are not suitable for this purpose, because they lack the necessary information of magnitude.

Measurement theory and statistics

Ensuring your data are fit for certain statistical tests is not the same as adhering to measurement theoretical principles (Figure 1). Although the goal of measurement theory is to make sure we extract meaningful inferences about the physical world, statistics make assumptions about the distributional properties of values to ensure that a particular test or model is producing sensible parameter estimates. Normality of residuals and linearity between variables are examples of common assumptions in many statistical approaches. When data do not meet these properties, it is common to apply various transformations in order to approximate these expectations. However, data transformations can also break the link between the value and the aspect of reality that the value represents, especially nonlinear transformations that are commonly used, such as arcsine and logarithmic transformations.

There are situations in which meeting the assumptions of statistical models may require manipulating the data in ways that do not align with measurement theoretical principles. This potential conflict between measurement theory and statistics can be a challenge in any quantitative field, including paleobiology. However, one of the key goals of statistical analysis is to make meaningful statements. 'Meaningful statistics' [4,18] produce results that are invariant to **allowed transformations** of the measurements, given their scale type. We therefore urge caution in transforming data to fit statistical models in ways that may violate measurement theoretical principles. Both measurement theory and statistics are crucial in ensuring that we draw meaningful inferences based on our data.

Measurement theoretical challenges in paleobiology

The fossil record can provide unique data that capture ecological and evolutionary processes and patterns across time intervals that are inaccessible to researchers studying only modern taxa. Distinct and significant contributions from paleobiology to the development of evolutionary biology include insights on the tempo and mode of phenotypic evolution within lineages (e.g., [19–23]), speciation and extinction dynamics of clades (e.g., [24–28]), phylogenetic inference (e.g., [29–32]), biogeographic patterns (e.g., [33–36]), and patterns of morphological disparity (e.g., [37–41]). However, a challenge for some types of studies within these topics, and in the quantitative study of the fossil record more generally, is that many of the statistical approaches used assume ratio-scaled variables, whereas parts of the data often belong to nominal and ordinal scale types. This mixture of scale types can create challenges for conducting meaningful statistics and thus robust inference.

Beware of mixing scale types

Combining and jointly analyzing measurements belonging to different scale types in a meaningful way is a challenge in all scientific disciplines, including paleobiology. We use analyses of disparity – the study of the evolution of variation in anatomy, function, and ecology of taxa and clades [38,41,42] – merely to exemplify some of these challenges, because disparity is widely studied in paleobiology. Studies of disparity can be conducted on measurements belonging to any scale type but often rely on character matrices where nominal and ordinal scale types dominate (e.g., [40,41,43–51]). There are good reasons for this, because many characters that are important for describing differences and similarities among taxa are nominal or ordinal in nature. It is also perfectly reasonable to want to include as many characters as possible in analyses where the goal is to describe the variation in morphology at the level of clades. However, producing meaningful statements when jointly analyzing data containing measurements of various scale types is far from straightforward.

Many disparity analyses start with the computation of a distance matrix based on a character matrix, the latter representing a summary of the character states of different traits in the taxa under study. The distance matrix contains the pairwise distances in trait space between all taxa in the study. However, distances between different traits and the character states across traits may not be directly comparable, because measurements on nominal and ordinal scale types do not possess the property of magnitude. Furthermore, distance metrics are commonly based on various transformations of the raw distances (e.g., [37,52–54]), which can further obscure the link between the measurements and the reality they are intended to represent. Disparity analyses also frequently make use of principal component analysis (PCA) and principal coordinate analysis (PCoA) to calculate variances and covariances among variables based on distance matrices. These ordination techniques are often used in paleobiology, for example, to reduce the dimensionality of large datasets in order to analyze and visualize the main axes of variation or dissimilarity (e.g., [55]). However, it is important to note that the outcome of PCA and PCoA (e.g., the numbers of the principal components), and consequently the meaningfulness of the results of these approaches, depends on whether the variables that are part of the analysis belong to a scale type where magnitudes between the variables are directly interpretable. This is often not the case if the data consist of a mix of scale types. Finally, some traits may not have been measured or may be absent in some taxa, and how to treat missing data in disparity analyses is frequently discussed (see next paragraph). How to deal with missing data is also a measurement theoretical challenge, because it is unclear how to put a value on something that either cannot be observed or could be observed but has not been measured. Thus, there are several potential measurement theoretical challenges in the study of disparity that may even render qualitative conclusions incorrect (Box 2).

The disparity literature is rich in methodological discussions that are related to but not the same as the measurement theory challenges mentioned above, such as how to treat missing data and its effects (e.g., [54,56–59]), how to incorporate phylogeny (e.g., [60]), or under which circumstances categorical or continuous trait data tell the same or different story (e.g., [48,61–63]). The effect of different data types and methodological assumptions on disparity metrics is important, but this is only one aspect of measurement theory and is something fundamentally different from ensuring the link between the measurements and the aspects of reality they represent is maintained to reach meaningful conclusions. Although being cognizant of the discrete/continuous distinction can prevent some violations of measurement theoretical principles, this distinction is often too broad to differentiate between scale types. For example, a recent study discussing best practice guidelines for disparity studies [42] lists the number of digits (ratio scale), the presence of webbed feet (nominal scale), and long/short wingspan (ordinal scale) as all under the discrete character mantle, but these traits clearly have different measurement theoretical properties, and

Box 2. Measurement theoretical challenges in disparity analyses

Disparity – the study of the structure and dynamics of morphological and ecological diversity, over time, across regions, and among taxonomic groups (e.g., [37,38,41,42,54]) – is a large field of inquiry in paleobiology. A measurement theoretical violation that often occurs in this subfield is applying the same statistical techniques to measurements belonging to different scale types, making inferences assuming all the measurements have the same properties. This practice can easily produce results that are difficult to interpret.

Assume we want to compare the disparity of the trilobites in Figure I in Box 1 over time (Figure II). Species A and species B are present in the first interval, whereas species B and C are present in the second. A simple disparity analysis to test if morphological diversity has increased or decreased through time can consist of calculating distance matrices and use the mean pairwise difference (mpd) as the disparity metric [54]. The encoding of a measurement belonging to an ordinal scale is arbitrary so long as it sustains the order; for example, the lateral glabellar furrows can be encoded as absent, partially developed, or present using either [0, 1, 2] or [2, 4, 8], or something else equally arbitrary. However, when calculating mean pairwise distances, the choice of encoding of [0, 1, 2] or [2, 4, 8] reverses the conclusion of this toy example, rendering the conclusions meaningless (Figure II). Thus, treating different scale types as equivalent and combining measurements belonging to different scale types in the same analysis may lead to spurious inferences, possibly more so in larger and more complex analyses.

Measurement theory may shed light on recurring themes and issues within disparity analyses, such as the degree to which metrics are 'Euclidean' [54], why different metrics can give different result (e.g., [53,71–73]), when to use ordination or not, and how to combine different types of traits. Disparity research may be strengthened and expanded by incorporating measurement theoretical aspects, not necessarily by producing obvious answers but by providing a foundation for more productive questions to be posed.
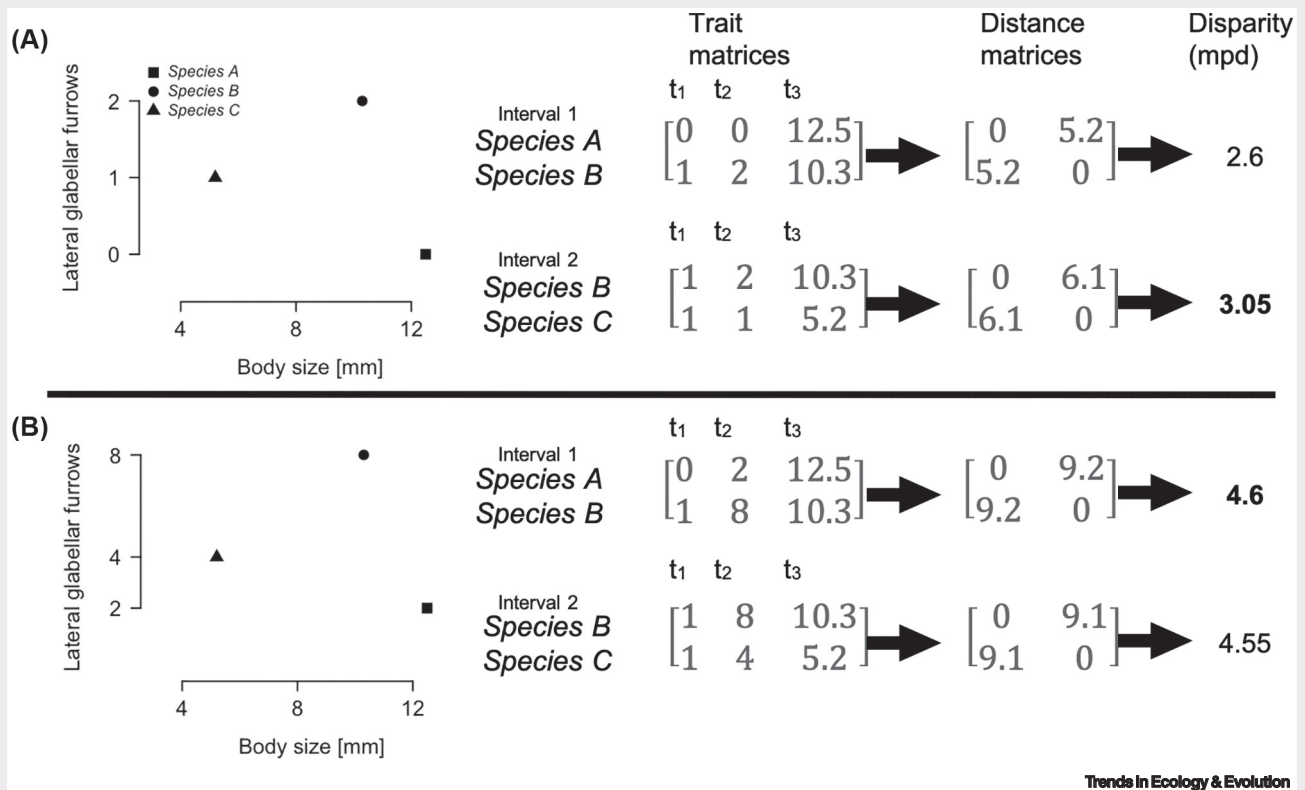

Figure II. Qualitative conclusions of disparity analyses can be contingent on encoding of traits. In both examples, the trilobite traits from Box 1 have been coded following the scale types in that box. In example (A), the lateral glabellar furrows of the trilobites have been encoded as absent = 0, partially developed = 1, and present = 2. Using the simple trait matrices and calculating disparity as mean pairwise distance leads to the conclusion that the first interval (species A + species B) is less diverse (mpd = 2.6) than the second interval (species B + species C, mpd = 3.05). In example (B), coding the furrows as absent = 2, partially developed = 4, and present = 8 leads to the opposite conclusion, despite both codings being equally valid for ordinal scale type measurements. This small example illustrates the impact of the coding scheme used to represent trait data on the conclusions when performing disparity analyses that do not respect the scale type. It also highlights the potential pitfalls in analyzing measurements from different scales assuming they hold the same amount of information.

analyzing them together can easily produce results that are not meaningful (Box 1). Thus, applying scale definitions from measurement theory can potentially aid in clearing a path forward by expanding the discrete/categorical versus continuous data dichotomy to ensure more meaningful results in the field.

Instead of jointly analyzing measurements belonging to different scale types using statistical machinery assuming ratio scale variables, a nonperfect but alternative approach is to transform the measurements so they belong to the same scale type. Although it is difficult to give a 'weak' scale type the same level of information as a 'stronger' one, it is usually feasible to reduce the information content of a measure on a 'stronger' scale type to a 'weaker' scale type. That is, transforming the ordinal categories of 'small,' 'medium,' and 'large' to meaningful ratio scale values in centimeters or millimeters is usually impossible, but discretizing the continuous variable into categories such as small, medium, and large is possible. Note, however, that collapsing measurements belonging to a 'stronger' scale type into a 'weaker' scale type reduces the information content in our original measurement.

How we choose to discretize ratio scale variables can also be somewhat arbitrary, however, and may potentially affect downstream analyses. In phylogenetic inference, for example, gap coding and its variants (gap weighting [64], step-matrix gap weighting [65]) are often used to discretize continuous variables for phylogenetic inference [66,67]. In its original formulation, trait means or medians were ranked from smallest to largest, and any gaps between successively ranked groups greater than some value (e.g., two within-population standard deviations [67]) served as boundaries between character state values. The researcher chooses the size of a gap and thus the number of character states. Although these gap sizes could possibly be guided by some biological principles, recent studies using simulation approaches suggest that increasing the number of character states can lead to well-resolved trees, but not necessarily those that reflect evolutionary history [68].

Instead of weakening stronger scale types, an alternative approach to dealing with mixed scale types is to analyze measures belonging to different scale types separately. For instance, if a character matrix consists of nominal, ordinal, and ratio scale variables, disparity can be computed individually for each scale type. This approach could allow more meaningful inference of changes in disparity across different types of traits, because it would at least not demand comparability among traits belonging to different scale types. Traits are usually measured the way they are for a reason. For example, nominal traits may represent the presence or absence of structures, which may be related more to the evolution of new traits in a clade than to the continuous variation in a trait captured by measures on a ratio scale. Although measurements belonging to the same scale type may not necessarily be directly comparable, separating analyses based on trait type may potentially provide insights on ecological and evolutionary patterns in qualitatively different kinds of traits.

Concluding remarks

Every scientist is essentially a practitioner of measurement theory, because scientists constantly make conscious or unconscious decisions about how to measure the particular aspect of reality they are studying. By using representational measurement theory, we can enhance our decision-making process and increase the likelihood of obtaining meaningful results when analyzing data and exploring ecological and evolutionary patterns and processes. To help avoid the most significant theoretical measurement mistakes, we have compiled a list of dos and don'ts in measurement theory (Box 3). However, it is essential to note that measurement theory is not an exact science in the sense that even measures within the same scale type may differ in their

Outstanding questions

How can we put a value on a character that is unobserved? Missing data are common in paleobiological analyses. Putting values on something that has not been measured is challenging and can affect the meaningfulness of downstream analyses.

How can datasets containing measures of different scale types be meaningfully analyzed? Measures of fossil taxa commonly belong to different scale types. Meaningful inference of joint analyses of mixed scale types is a measurement theoretical challenge.

What should we do when transformations of our data are needed to fulfill assumptions of a statistical model, but these transformations detach our measurements from the aspect of reality they originally represented? Models often vary in how robust they are to violations of underlying assumptions, but this needs investigation on a case-by-case basis.

Are scale types waiting to be discovered and defined due to the nature of paleobiological data? Applying measurement theory in the analysis of fossil data may uncover new measurement theoretical discoveries.

Box 3. Measurement theory's dos and don'ts

Identifying the scale type of a measurement is imperative and is step 1 in any application of measurement theory. The following four questions categorize measurements into the scale types of nominal, ordinal, interval, and ratio, from lower to higher on the hierarchy (see Box 1), and should be asked in ascending order:

1. Could the values be replaced with characters/symbols and not lose any information?
   a. Yes (go to 2), the data belong to a nominal or ordinal scale.
   b. No (go to 3).
2. Does the order of the characters carry any meaning?
   a. Yes, the data belong to an ordinal scale.
   b. No, the data belong to a nominal scale.
3. Do the measurements have units in which increments are placed at equal distance from one another? If yes, the data belong to either an interval or ratio scale (go to 4).
4. Does a value of 0 mean there is 'nothing' of the entity in question?
   a. Yes, the data belong to a ratio scale type.
   b. No, the data belong to an interval scale.

To avoid violating basic measurement theoretical principles, remember these points (the points are on a nominal scale):

- Means and variances of measures on a nominal or ordinal scale type should never be calculated. Calculating means and variances assumes distances between numbers carry meaning, a property that is not part of measures on a nominal and ordinal scale.
- Medians and modes can be calculated on ordinal scale variables because these are measures of central tendency in an ordered set of measures and do not assume meaningful distances between values.
- Numbers and frequencies of nominal and ordinal measures in a sample can be meaningful (e.g., the number of females in a population was 35, which represents 60% of the population).
- Logarithmic transformation should be applied only to measures on a ratio scale. Logarithmic transformation assumes that 0 means 'nothing of the entity measured.' Statements such as 'twice as large' and '20% less' are only meaningful if there is a true 0.
- Nonlinear transformations of data are often warranted (given the theoretical context of a study) but have consequences. Any nonlinear transformation of measurements (e.g., arcsine, logarithmic transformation, square root) changes the original link between the measurements and the reality they (initially) represented. Fitting a model to data before and after a nonlinear transformation means investigating different hypotheses.
- Applying the same statistical technique on variables belonging to different scale types assumes all the variables have the same properties. This is the same as making unjustified assumptions about the data.

---

accuracy and precision in reflecting the empirical world [1]. Paleobiology might also pose particular challenges to measurement theory, which could require new developments within measurement theory itself [1]. Additionally, classifying measurements into established scale types can be a challenging task (see Outstanding questions). Despite these pitfalls, applying measurement theory to studies of the fossil record provides researchers with assurance that their results are meaningful and have the potential to reveal new insights into the natural world.

Declaration of interests

The authors have no interests to declare.

References

1. Stevens, S.S. (1946) On the theory of scales of measurement. *Science* 103, 677–680
2. Krantz, D.H. *et al.* (1971) *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*, Academic Press
3. Stahl, W.R. (1962) Similarity and dimensional methods in biology. *Science* 137, 205–212
4. Houle, D. *et al.* (2011) Measurement and meaning in biology. *Q. Rev. Biol.* 86, 3–34

5. Rosen, R. (1978) *Fundamentals of Measurement and Representation of Natural Systems*, Elsevier
6. Smith, P.G.R. and Theberge, J.B. (1987) Evaluating natural areas using multiple criteria: theory and practice. *Environ. Manag.* 11, 447–460
7. Wolman, A.G. (2006) Measurement and meaningfulness in conservation science. *Conserv. Biol.* 20, 1626–1634
8. Voje, K.L. *et al.* (2020) Revisiting a landmark study system: no evidence for a punctuated mode of evolution in *Metrarabdotos*. *Am. Nat.* 195, 899–917
9. Huttegger, S.M. and Mitteroecker, P. (2011) Invariance and meaningfulness in phenotype spaces. *Evol. Biol.* 38, 335–351
10. Gould, S.J. (1980) The promise of paleobiology as a nomothetic, evolutionary discipline. *Paleobiology* 6, 96–118
11. Sepkoski, D. (2012) *Rereading the Fossil Record: The Growth of Paleobiology and an Evolutionary Discipline*, University of Chicago Press
12. Sepkoski, D. and Ruse, M. (2009) *The Paleobiological Revolution: Essays on the Growth of Modern Paleontology*, University of Chicago Press
13. Smith, J.M. (1984) Evolution: palaeontology at the high table. *Nature* 309, 401–402
14. Suppes, P. *et al.* (1989) *Foundations of Measurement, Volume II: Geometrical, Threshold, and Probabilistic Representations*, Academic Press
15. Luce, R.D. *et al.* (1990) *Foundations of Measurement, Volume III. Representation, Axiomatization, and Invariance*, Academic Press
16. Stockdale, M.T. and Benton, M.J. (2021) Environmental drivers of body size evolution in crocodile-line archosaurs. *Commun. Biol.* 4, 38
17. Benson, R.B.J. *et al.* (2022) Reconstructed evolutionary patterns for crocodile-line archosaurs demonstrate impact of failure to log-transform body size data. *Commun. Biol.* 5, 171
18. Adams, E.W. *et al.* (1965) A theory of appropriate statistics. *Psychometrika* 30, 99–127
19. Simpson, G.G. (1944) *Tempo and Mode in Evolution*, Columbia University Press
20. Eldredge, N. and Gould, S.J. (1972) Punctuated equilibria: An alternative to phyletic gradualism. In *Models in Paleobiology* (Schopf, T.J.M., ed.), pp. 82–115, Freeman Cooper
21. Hunt, G. (2012) Measuring rates of phenotypic evolution and the inseparability of tempo and mode. *Paleobiology* 38, 351–373
22. Hunt, G. (2007) The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. *Proc. Natl. Acad. Sci. U. S. A.* 104, 18404–18408
23. Voje, K.L. (2016) Tempo does not correlate with mode in the fossil record. *Evolution* 70, 2678–2689
24. Raup, D.M. and Jr, J.J.S. (1982) Mass extinctions in the marine fossil record. *Science* 215, 1501–1503
25. Alroy, J. *et al.* (2001) Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proc. Natl. Acad. Sci. U. S. A.* 98, 6261–6266
26. Bambach, R.K. (2006) Phanerozoic biodiversity mass extinctions. *Annu. Rev. Earth Planet. Sci.* 34, 127–155
27. Kocsis, Á.T. *et al.* (2019) The r package divDyn for quantifying diversity dynamics using fossil sampling data. *Methods Ecol. Evol.* 10, 735–743
28. Foote, M. (2023) Diversity-dependent diversification in the history of marine animals. *Am. Nat.* 201, 680–693
29. Donoghue, M.J. *et al.* (1989) The importance of fossils in phylogeny reconstruction. *Annu. Rev. Ecol. Syst.* 20, 431–460
30. Gould, S.J. *et al.* (1977) The shape of evolution: a comparison of real and random clades. *Paleobiology* 3, 23–40
31. Koch, N.M. and Parry, L.A. (2020) Death is on our side: paleontological data drastically modify phylogenetic hypotheses. *Syst. Biol.* 69, 1052–1067
32. Ronquist, F. *et al.* (2012) A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. *Syst. Biol.* 61, 973–999
33. Jablonski, D. *et al.* (2006) Out of the tropics: evolutionary dynamics of the latitudinal diversity gradient. *Science* 314, 102–106
34. Valentine, J.W. and Moores, E.M. (1970) Plate-tectonic regulation of faunal diversity and sea level: a model. *Nature* 228, 657–659
35. Rosen, B.R. (1988) From fossils to earth history: applied historical biogeography. In *Analytical Biogeography* (Myers, A.A. and Giller, P.S., eds), pp. 437–481, Springer
36. Lieberman, B.S. (2003) Paleobiogeography: the relevance of fossils to biogeography. *Ecol. Evol. Syst.* 34, 51–69
37. Wills, M.A. *et al.* (1994) Disparity as an evolutionary index: a comparison of Cambrian and Recent arthropods. *Paleobiology* 20, 93–130
38. Foote, M. (1999) Morphological diversity in the evolutionary radiation of Paleozoic and Post-Paleozoic crinoids. *Paleobiology* 25, 1–115
39. Liow, L.H. (2007) Lineages with long durations are old and morphologically average: an analysis using multiple datasets. *Evolution* 61, 885–901
40. Brusatte, S.L. *et al.* (2008) Superiority, competition, and opportunism in the evolutionary radiation of dinosaurs. *Science* 321, 1485–1488
41. Hughes, M. *et al.* (2013) Clades reach highest morphological disparity early in their evolution. *Proc. Natl Acad. Sci.* 110, 13875–13879
42. Guillerme, T. *et al.* (2020) Disparities in the analysis of morphological disparity. *Biol. Lett.* 16, 20200199
43. Young, M.T. *et al.* (2010) The evolution of Metriorhynchoidea (mesoeucrocodylia, thalattosuchia): an integrated approach using geometric morphometrics, analysis of disparity, and biomechanics. *Zool. J. Linnean Soc.* 158, 801–859
44. Prentice, K.C. *et al.* (2011) Evolution of morphological disparity in pterosaurs. *J. Syst. Palaeontol.* 9, 337–353
45. Thorne, P.M. *et al.* (2011) Resetting the evolution of marine reptiles at the Triassic-Jurassic boundary. *Proc. Natl Acad. Sci.* 108, 8339–8344
46. Bapst, D.W. *et al.* (2012) Graptoloid diversity and disparity became decoupled during the Ordovician mass extinction. *Proc. Natl Acad. Sci.* 109, 3428–3433
47. Ruta, M. *et al.* (2006) Evolutionary patterns in early tetrapods. I. Rapid initial diversification followed by decrease in rates of character change. *Proc. R. Soc. B Biol. Sci.* 273, 2107–2111
48. Hetherington, A.J. *et al.* (2015) Do cladistic and morphometric data capture common patterns of morphological disparity? *Palaeontology* 58, 393–399
49. Hopkins, M.J. and Smith, A.B. (2015) Dynamic evolutionary change in post-Paleozoic echinoids and the importance of scale when interpreting changes in rates of evolution. *Proc. Natl Acad. Sci.* 112, 3758–3763
50. Sutherland, J.T.F. *et al.* (2019) Does exceptional preservation distort our view of disparity in the fossil record? *Proc. R. Soc. B* 286, 20190091
51. Bolet, A. *et al.* (2022) The Jurassic rise of squamates as supported by lepidosaur disparity and evolutionary rates. *Elife* 11, e66511
52. Gower, J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics* 27, 857
53. Hopkins, M.J. and John, K.S. (2018) A new family of dissimilarity metrics for discrete character matrices that include inapplicable characters and its importance for disparity studies. *Proc. R. Soc. B* 285, 20181784
54. Lloyd, G.T. (2016) Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, progress, and future directions. *Biol. J. Linn. Soc.* 118, 131–151
55. Pimiento, C. *et al.* (2020) Functional diversity of marine megafauna in the Anthropocene. *Sci. Adv.* 6, eaay7650
56. Clavel, J. *et al.* (2014) Missing data estimation in morphometrics: how much is too much? *Syst. Biol.* 63, 203–218
57. Deline, B. and Thomka, J.R. (2017) The role of preservation on the quantification of morphology and patterns of disparity within Paleozoic echinoderms. *J. Paleontol.* 91, 618–632
58. Smith, A.J. *et al.* (2014) Joined at the hip: linked characters and the problem of missing data in studies of disparity. *Evolution* 68, 2386–2400
59. Gerber, S. (2019) Use and misuse of discrete character data for morphospace and disparity analyses. *Palaeontology* 62, 305–319
60. Brusatte, S.L. *et al.* (2011) Phylogenetic corrections for morphological disparity analysis: new methodology and case studies. *Paleobiology* 37, 1–22

61. Schaeffer, J. *et al.* (2020) Morphological disparity in theropod jaws: comparing discrete characters and geometric morphometrics. *Palaeontology* 63, 283–299

62. Ferrón, H.G. *et al.* (2020) Categorical versus geometric morphometric approaches to characterizing the evolution of morphological disparity in Osteostraci (Vertebrata, stem Gnathostomata). *Palaeontology* 63, 717–732

63. Villier, L. and Eble, G.J. (2004) Assessing the robustness of disparity estimates: the impact of morphometric scheme, temporal scale, and taxonomic level in spatangoid echinoids. *Paleobiology* 30, 652–665

64. Thiele, K. (1993) The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics* 9, 275–304

65. Wiens, J.J. (2001) Character analysis in morphological phylogenetics: problems and solutions. *Syst. Biol.* 50, 689–699

66. Riska, B. (1979) Character variability and evolutionary rate in Menidia. *Evolution* 33, 1001–1004

67. Johnson, M.S. and Mickevich, M.F. (1977) Variability and evolutionary rates of characters. *Evolution* 31, 642–648

68. Bardin, J. *et al.* (2014) Increasing the number of discrete character states for continuous characters generates well-resolved trees that do not reflect phylogeny. *Integr. Zool.* 9, 531–541

69. Monks, N. and Palmer, P. (2002) *Ammonites*, Smithsonian Books

70. Moriya, K. (2015) Isotope signature of ammonoid shells. In *Ammonoid Paleobiology: From Anatomy to Ecology* (Klug, C. *et al.*, eds), pp. 793–836, Springer

71. Hopkins, M.J. (2022) Single disparity indices can be misleading: comment on. *Lethaia* 55, 1–3

72. Suárez, M.G. and Esteve, J. (2021) Morphological diversity and disparity in trilobite cephala and the evolution of trilobite enrolment throughout the Palaeozoic. *Lethaia* 54, 752–761

73. Wills, M.A. (2001) Morphological disparity: a primer. In *Fossils, Phylogeny, and Form, an Analytical Approach* (Adrain, J.M. *et al.*, eds), pp. 55–145, Kluwer Academic/Plenum Publishers