Measuring evolutionary independence: A pragmatic approach to species classification [accepted manuscript]

Stijn Conix, Centre for Logic and Philosophy of Science, Institute of Philosophy, KU Leuven.

Abstract

After decades of debates about species concepts, there is broad agreement that species are evolving lineages. However, species classification is still in a state of disorder: different methods of delimitation lead to competing outcomes for the same organisms, and the groups recognised as species are of widely different kinds. This paper considers whether this problem can be resolved by developing a unitary scale for evolutionary independence. Such a scale would show clearly when groups are comparable and allow taxonomists to choose a conventional threshold of independence for species status. Existing measurement approaches to species delimitation are typically shot down by what I call the heterogeneity objection, according to which independently evolving groups are too heterogeneous to be captured by a single scale. I draw a parallel with the measurement of temperature to argue that this objection does not provide sufficient reasons to abandon the measurement approach, and that such an approach may even help to make the vague notion of evolutionary independence more precise.

Keywords: evolutionary independence; evolving lineages; temperature measurement; species; species delimitation; taxonomic disorder;

Resolving taxonomic disorder

In a recent comment in *Nature,* Stephen Garnett and Les Christidis (2017) draw attention to what they call 'taxonomic anarchy': groups recognised as species are of widely different kinds, species classifications are unstable, and experts often disagree with the classifications that fellow experts come up with. This disorder might seem surprising, given broad agreement that species are evolving lineages (De Queiroz 1998; Zachos 2016; Barker 2019). However, this agreement is merely superficial, as the notion of evolutionary lineage is understood and operationalised in a large variety of ways. Because different operationalisations often yield different results for the same groups of organisms, even taxonomists with the same theoretical view on species regularly come up with competing classifications (Willis 2017; Zachos 2018; Conix 2018).

This state of disorder has direct consequences for scientific research that uses species as its currency and thus assumes that species are comparable. For example, the incomparability of species affects measures of biodiversity that rely on species richness, diversification analyses, and different kinds of macro-ecological research (Isaac et al. 2004; Faurby et al. 2016; Zachos 2018). The consequences of taxonomic disorder also extend beyond science to domains such as conservation legislation and the regulation of trade in endangered species, which extensively rely on the outcomes of taxonomy. Particularly noteworthy are the implications of taxonomic disorder for biodiversity conservation. The lack of a reliable and easily accessible inventory of biodiversity forms a major bottleneck in addressing the ongoing mass extinction, and taxonomic disorder is often cited as one important factor standing in the way of such an inventory: it impedes conservation biology and effective conservation legislation, discourages funding organizations from investing in taxonomy, leads to wasting resources on competing species lists, and makes taxonomic information difficult to access for non-specialists (Mace 2004; Khuroo et al. 2007; Garnett and Christidis 2017).

It is clear that a solution for taxonomic disorder is desirable, and it should be no surprise that users of taxonomy regularly call for such a solution (Godfray 2002; Isaac et al. 2004). One potential solution that recurs regularly in these debates is the proposal to standardise species delimitation (Conix 2019). The idea is that if all species are recognised on the basis of similar principles or methods, the resulting species would be comparable and taxonomic disorder greatly reduced. Such standardisation could start from the consensus that species are evolving lineages. However, this solution is typically rejected on the grounds of what I will call the heterogeneity objection: there is no single standard that can capture the vast variety of evolving lineages in the organic world.

The main aim of this paper is to argue that the heterogeneity objection does not provide good reasons to abandon attempts to standardise species delimitation. More positively, I argue that developing measurement procedures and a scale of evolutionary independence – a notion frequently used to characterise evolutionary lineages – is a promising strategy for standardisation. Note that the measurement approach defended here is not an alternative conception of what species *are*, nor is it a claim about what evolutionary lineages are like. Rather, the measurement approach is a strategy to tackle the problem of taxonomic disorder, and my claim is that it is a strategy worth pursuing. More precisely, the aim of this paper is to show that, contrary to popular opinion, measurement of evolutionary independence could be successful despite the heterogeneity of evolving lineages.

To do this, I first introduce evolutionary independence and show how standardisation could proceed by measuring this property ("Evolutionary independence, measurement and the Tobias criterion"). I then discuss the heterogeneity objection and argue that it is best interpreted as the claim that evolutionary independence cannot be measured because it is theoretically vague, heterogeneous, and operationalised in divergent ways ("The heterogeneity objection"). In response to this objection, I draw a parallel with the measurement of temperature. This parallel shows that the heterogeneity objection does not provide good reasons to abandon the measurement of evolutionary independence ("Measurement despite heterogeneity: The case of temperature"), and that measurement may even help to make the notion of evolutionary independence more precise ("Measurement as the way out"). Finally, I further qualify this point as an appeal for pragmatism about species classification ("A plea for pragmatism"), and conclude by observing the fit between this pragmatic measurement approach and recent developments in species delimitation.

Evolutionary independence, measurement and the Tobias criterion

Despite longstanding and ongoing debates about the nature of species, there is wide agreement at least that species are evolving lineages (e.g. De Queiroz 1998; Zachos 2016; Barker 2019). Such lineages are ancestor-descendant sequences of populations that take part in, and are subject to, the same evolutionary processes. Because the populations and organisms that make up evolving lineages are connected to each other by evolutionary processes, these lineages are often called *units of evolution* that have a *shared evolutionary fate* distinct from that of other lineages. These units are typically seen as active in the sense that they take part in evolutionary processes. In this respect evolutionary lineages are different from lineages defined as monophyletic clades, which are merely the passive patterns that result from these processes (Baum 2009).

Evolutionary lineages play a central role in current systematic research. For example, they are the target of the popular statistical model-based approaches to species delimitation. These model-based approaches use genetic or genomic datasets along with a set of assumptions about evolutionary processes to infer the history of diversification and species boundaries in the dataset (Edwards 2009; Camargo and Sites 2013; Carstens et al. 2013). Most popular among these model-based approaches are methods using the multispecies coalescent model (Yang and Rannala 2010; Fujita et al. 2012). This model, which connects population genetics and phylogenetics, estimates species trees while taking into account some of the population-level processes that cause discordance between species trees

and gene trees. This allows systematists to identify independently evolving lineages despite such discordance.

Despite the central role of evolutionary lineages in systematic research, the notion has been cashed out in different ways (Barker 2019). Many of these define evolutionary lineages in terms of particular processes or causes, such as gene flow or shared selection pressures, that are responsible for the cohesion of these lineages. More useful for the purposes of standardisation are Kevin De Queiroz' (1998) General Lineage Concept (GLC), plus broadly similar views by Mayden (1997), Wiley (1978), and Simpson (1961). Their so-called evolutionary conceptions of species do not privilege particular causes that need to be involved in the cohesion of lineages for them to count as evolving lineages. Instead, they recognize that these causes can be different for each lineage, and emphasize that despite this variation all species are similar in that they are *independently evolving* units. The broad applicability of these views, as well as their popularity among taxonomists, make them an obvious starting point for reducing taxonomic disorder.

However, while the broadness of these evolutionary views of species allows them to capture the different ways in which groups of populations can be independently evolving lineages, it lies at the heart of their main weakness too. Phrases such as 'independent evolution', 'a common evolutionary fate' (Templeton 1992, p. 160), or 'separate identities' (Willmann and Meier 2000, p. 116) are metaphors, and critics argue that evolutionary conceptions of species are too vague to go beyond this metaphorical level (Barker 2019, pp. 9–11). In the absence of precise criteria for what constitutes independent evolution, unifying species under the label of independently evolving lineages may contribute more to taxonomic disorder than it does to resolving it.

Three related sources of disorder are worth distinguishing here: evolutionary independence is continuous, independently evolving groups are not mutually exclusive, and evolutionary independence comes in many different shapes. First, while evolutionary independence was traditionally understood as a discrete property, it is now generally recognised to be a continuous property. For example, the evolutionary independence of two diverging lineages during speciation typically increases gradually as gene flow decreases and they acquire different characters (Roux et al. 2016). It follows that there is no obvious point on the continuum that distinguishes the species level from more and less exclusive groups, and methods of species delimitation have to select a partially

arbitrary threshold of independent evolution to distinguish lineages that deserve species status from lineages that do not.[1]

Second, genomic studies suggest that due to processes such as hybridization, lateral gene transfer and introgression, it is not uncommon for groups of organisms to evolve together for some traits while being independent for others (Mallet et al. 2015; Haber 2019). In addition, evolving lineages often contain smaller lineages and are nested in larger lineages. This means that populations are typically simultaneously part of multiple nested or overlapping evolving lineages. Because current practices of species classification assume that each population can only be part of one species, taxonomists have to decide which of these to recognise as species.

Finally, different lineages evolve independently in widely different ways. Some independently evolving groups are reproductively isolated, others are subject to similar selective pressures, and others evolve independently in still other ways. To delimit species, therefore, taxonomists have to use a wide array of operationalisations suitable to the diversity of evolving lineages they encounter. When applied to the same groups, however, these operationalisations often point in different directions (Willis 2017; Zachos 2018). For example, a group may share the same selection pressures but not be reproductively isolated, or vice versa. Thus, taxonomists engaged in species delimitation sometimes have to choose between multiple divergent operationalisations.

The result is taxonomic disorder: even taxonomists who work with the same organisms can come up with different classifications depending on how they operationalise evolutionary independence, where they set the threshold for species status, and which of several nested or partially overlapping lineages they recognise. To standardise species classification and resolve taxonomic disorder, then, more is needed than agreement on the view that species are independently evolving lineages. In addition, we would need a unitary scale that allows us to evaluate and compare lineages that are evolving independently in different ways. With such a scale in hand, it would be possible to set a threshold of evolutionary independence required for species status, and recognise all groups that meet this threshold. This would yield species that are comparable at least in the interesting sense that they are evolving independently to the same degree.

[1] Note that by this continuum I do not mean a sequence of different indicators of evolutionary independence (e.g. diagnosability, reproductive isolation, etc.) such as on De Queiroz' famous diagram of diverging lineages (De Queiroz, 1998, p. 64 fig. 5.4). De Queiroz' diagram is problematic because these indicators often occur in different orders, shapes, and combinations. Instead, I propose thinking of evolutionary independence as a quantity (perhaps ordinal) that any group of populations has to some degree.

Given that this solution, if possible, would effectively resolve taxonomic disorder, it should not be surprising that various researchers have recently proposed to standardise species delimitation by means of what critics have termed 'yardstick approaches' (Halley et al. 2017, p. 390).[2] These are standardised procedures that can be seen as assigning a certain degree of evolutionary independence to groups of organisms (Hey and Pinho 2012; Galtier 2019). To illustrate the measurement approach to species delimitation, the remainder of this section will briefly present one such yardstick method, namely, the Tobias criterion (Tobias et al. 2010).[3] Note that I choose the Tobias criterion to illustrate the measurement approach because it is relatively simple, has received wide attention from both taxonomists and users of taxonomy (Remsen 2016; Burfield et al. 2017; Halley et al. 2017), and has been adopted in practice by the prominent organization Birdlife International for their species checklist (Hoyo and Collar 2014). It is not my aim to defend the Tobias criterion as the best method of species delimitation or even the most promising approach to measuring evolutionary independence.[4]

The Tobias criterion is a standardised method for delimiting bird species on the basis of phenotypic divergence. Simply put, it assigns a score between one and four to three main kinds of traits (voice, morphology and biometrics), and a score of one or two to one subsidiary trait (ecological or behavioural divergence). In addition, a group gets a score between zero and three on the basis of its geographical relationships (ranging from allopatry to parapatry). These scores are then summed, and a group is recognised as a species if it scores seven or more. The degrees for each criterion are defined in terms of quantitative thresholds. These thresholds were set such that the approach recognises 95% of uncontroversial sympatric species at a score of seven. This way, the scores of a reference set of uncontroversial sympatric species form a yardstick to delimit new species or evaluate difficult cases like allopatric groups (Tobias et al. 2010, p. 731).

Note that species delimitation using the Tobias criterion has all the components by which metrologists and philosophers of measurement typically define measurement (Cartwright et al. 2011; Tal 2013;

[2] The connection between species delimitation and measurement has been made in philosophy too. More precisely, Conix (2018) draws an analogy between measurement and integrative methods of species delimitation. The claim in this paper is more radical, however, as I do not merely draw an analogy between one taxonomic method and measurement in the physical sciences, but instead propose to see species delimitation as a form of measurement.

[3] It is typically referred to as the Tobias criteria, referring to the multiple parameters in the model. I use 'criterion' here to emphasize that it integrates these parameters into a single score for species status.

[4] Indeed, it could be argued that a viable measurement approach should build on existing measures such as those of genetic divergence, population connectivity and the strength of isolation barriers (Coyne and Orr 1989; Palsbøll et al. 2007; Lowe and Allendorf 2010; Sobel et al. 2010). However, the Tobias criterion takes none of these into account.

Mari et al. 2017). Most obviously, there is a clearly defined set of objects (i.e. groups of organisms or populations), a target property (i.e. evolutionary independence) attributed to those objects, and a measurement instrument (i.e. the Tobias criterion). This measurement instrument is calibrated by means of a reference set of sympatric uncontroversial species that acts as a classical measurement standard. Less obviously, and in line with the consensus in recent epistemology of measurement, theory and background knowledge play a crucial role in the measurement process. Evolutionary independence is characterized theoretically as being on a unique trajectory through evolutionary space. The measurement instrument is designed in line with this theoretical characterization and further background knowledge. For example, it is known that indicators within one category of criteria are often causally related. For this reason, the system caps the number of points that a group can get within one category of criteria. This way, one species cannot get points both for, say, plumage pattern and plumage colour.

A measurement method like the Tobias criterion has the potential to reduce taxonomic disorder in various ways. First, different taxonomists using the same data would get the same classifications, as the procedure and thresholds for species delimitation are standardised. This mitigates the problem of what Schlick-Steiner et al. (2010, p. 421) call 'researcher bias' in taxonomy: species classifications can turn out differently depending on the subjective decisions of the taxonomist conducting the research. Second, the outcomes of taxonomy would be unlikely to be biased by the choice of one particular indicator of evolutionary independence, as the Tobias criterion takes into account a range of independent indicators and maps them on a single scale. Finally, recognised species would arguably be more comparable because major sources of incomparability, such as the use of only one indicator of evolutionary independence, are eliminated. It should also be mentioned that the method's ability to reduce disorder can be expected to increase over time, as the transparency of the method facilitates further improvements and tweaking of the measure (Burfield et al. 2017; Donegan 2018).

Of course, the Tobias criterion faces problems too. Indeed, these problems have proven substantial enough to prevent most taxonomists from adopting it (Rheindt et al. 2011; but see e.g. Balen et al. 2013; Schuchmann et al. 2016). However, most of these problems do not concern the measurement approach, but the particular way the Tobias criterion implements it. Most importantly, the Tobias criterion bypasses genetic data completely, thus ignoring much of the progress made in species delimitation over the past decade. In addition, the criterion has been criticised for the way it scores hybrid zones, its use of hard cut-offs for mostly continuous criteria, and the subjective evaluation of plumage pattern it requires (Remsen 2015, 2016; Donegan 2018). Finally, the system is obviously only

applicable to birds, and thus resolves only a small part of taxonomic disorder. As Donegan (2018), Tobias et al. (2010) and Burfield et al. (2017) argue, these problems can easily be mitigated. For example, additional scores for genetic data could be implemented, and spectrophotometry could be used to quantitatively assess plumage patterns and colours. The approach could also be extended to other taxa by adding further criteria, adapting the reference set accordingly, and setting new thresholds.

A more substantial problem is signalled by Remsen (2015), who points out that the sample of species in the reference set is phylogenetically biased as the species are almost exclusively taken from the order of *Passeriformes*. It seems dubious, then, to apply the system to other orders of birds too. As Remsen (2016, p. 112) puts it, 'who cannot appreciate that a scoring scheme derived almost completely from small diurnal passerines might not be appropriate for assessing species limits in, for example, petrels or owls?' While this problem may seem easy to resolve by broadening the reference set, it points to deeper problems that go to the core of the measurement approach: *how can a representative set be selected from the extreme variety of taxa throughout the tree of life, and will the resulting measure not simply churn out whatever kind of variety we decide to put in the reference set in the first place?* Remsen's objection questions whether it is in principle possible to create a unitary scale of evolutionary independence that is applicable across different groups. I turn to this crucial objection, which I call the heterogeneity objection, in the next section.

The Heterogeneity objection

The heterogeneity of independently evolving groups is far more extensive than suggested by Remsen's quote about different bird species. Most obviously, the notion applies to groups like *Homo sapiens*, which evolve independently from other lineages due to extensive gene flow within the species and little exchange of genes with other species. Such evolutionary independence can be contrasted with that of groups like Edith's checkerspot (*Euphydryas editha*), a butterfly species that retains high evolutionary cohesion despite a lack of regular gene flow between populations (Ehrlich and Raven 1969). Similarly, it is substantially different from the evolutionary independence of species in syngameons, which are collections of species that evolve independently despite substantial gene flow between them (Templeton 1992; Barker 2007). The evolutionary independence of all these cases is in turn radically different from that of *Erythranthe peregrina*, a polyploid species of monkeyflower from Scotland that is independent from its diploid parental species because it has a different number of chromosomes (Vallejo-Marín et al. 2015); a facultative parthenogenetic species such as the New

Zealand Mud snail (*Potamopyrgus antipodarum*); and groups that exchange nuclear genes but not mitochondrial ones, such as *Discoglossus jeanneae* and *galganoi* (García-París and Jockusch 1999).

Anyone familiar with recent philosophy of biology knows that this heterogeneous list of independently evolving lineages could be extended indefinitely. These examples are also complemented by growing insight into discordance between lineages at different hierarchical levels (Maddison 1997; Haber 2012, 2019). Most importantly, it is well known that discordance between gene trees and species trees due to phenomena such as incomplete lineage sorting, introgression, lateral gene transfer and hybridization is very common. This means that within any organism, some genes are more closely related to those of organisms in another species than to those of its conspecifics. This substantially complicates the notion of evolutionary independence, as it means that even within eukaryotes there is no one-on-one mapping between lineages at different hierarchical levels.

Various biologists have objected to the idea of a unitary scale of evolutionary independence on the grounds of this heterogeneity. For example, in his discussion of the Tobias criterion, Galtier (2019) asks 'how can a standard be defined when … speciation in nature can follow so many different routes?' Similarly, Halley et al. (2017, p. 390) remark that choosing a yardstick is 'biased by sampling error that stems from the vagaries of extinction and incomplete sampling'. More broadly, a similar objection seems to follow directly from theoretical species pluralism. While many pluralists accept that all species are lineages, they emphasise that there are different kinds of lineages that evolve independently in different senses. In his discussion of De Queiroz' GLC, for example, Marc Ereshefsky (2011, p. 75) writes that the catch-all concept of independently evolving lineage just 'masks the heterogeneity of the species category because what constitutes a lineage has multiple answers, and those answers vary according to which species concept [i.e. operationalisation] one adopts'. In other words, integrating different operationalisations of evolutionary independence under a single concept is not helpful because they pick out different interesting features from the organic world.

The picture that emerges from these objections can be captured in three related claims. First, the notion of evolutionary independence is theoretically vague. It is typically clarified by means of metaphors, and whenever it is developed in more precise terms it seems to take a different meaning in different contexts of use. Thus, rather than picking out a precise quality or quantity, it seems to capture a loose family resemblance cluster of properties. Second, and closely related to this, the concept is manifested in a heterogeneous collection of groups. While all these groups are characterised by some of the properties associated with the family resemblance cluster, there are no

core properties they all have in common. Third, and as a result of the previous two points, there are many competing operationalisations of evolutionary independence, and there is no principled way of choosing between them. Depending on which operationalisation is chosen, different groups are individuated as independently evolving lineages.

Taken together, these three points pose a seemingly insurmountable hurdle for the measurement of evolutionary independence. The problem is best represented as a sort of catch-22: The concept of evolutionary independence is theoretically vague. Because of this vagueness, it applies to a wide variety of groups but does not make them comparable in an interesting way. To make the concept useful, then, it needs to be made more precise. Biologists try exactly this when they operationalise the notion. However, these operationalisations are typically only useful for a small part of all independently evolving lineages and are often in conflict with each other. Because of this, the operationalisations also fail to make species comparable. To make the operationalisations of evolutionary independence useful for resolving taxonomic disorder, then, we need a way of arbitrating between them. However – and this brings us full circle – this is only possible if the theoretical notion itself is more precise. Thus, resolving the theoretical vagueness of evolutionary independence requires better operationalisations, and improving operationalisations requires a precise theory.

This heterogeneity fuelled catch-22 seems to render the project of developing a unitary scale of the quantity – and, indeed, calling it a quantity to begin with – misguided: lacking both a precise theory and broadly applicable operationalisations, it seems impossible to construct a unitary scale of evolutionary independence that allows us to compare groups as heterogeneous as the ones we find in the organic world.

Measurement despite heterogeneity: The case of temperature

The heterogeneity objection states that measurement of evolutionary independence is impossible because the notion is vague, applies to a heterogeneous collection of measurands (things being measured), and is operationalised in many different ways. Together, the objection states, these characteristics lead to a catch-22 that makes measurement impossible. It follows from this that biologists should not pursue the measurement of evolutionary independence. This section suggests that this conclusion is too quick. I argue that even if evolutionary independence is vague, heterogeneous and operationalised in multiple ways, it does not follow that measurement cannot

succeed. To do this, I show that the same three characteristics were once true for the physical notion of temperature. As measurement of temperature is highly successful and precise today, it follows that these three characteristics do not make the development of a useful measurement scale impossible in the way the heterogeneity objection suggests.

Before showing how the thee characteristics once applied to temperature, it is worth addressing the objection that temperature and evolutionary independence are simply too different for the analogy to work. There are at least four such differences. First, temperature is a physical quantity that applies to physical systems while evolutionary independence applies to groups of biological populations partaking in evolutionary processes. Second, temperature is defined on the molecular level, while evolutionary independence is defined on the macro-level of organisms and populations. Third, evolutionary independence is a purely theoretical term while temperature is both a theoretical term and a natural language term for describing day-to-day sensations. Finally, and perhaps most importantly, temperature is a precise quantity that does not depend on particular decisions made by those who measure it. That is, while measuring temperature requires choosing a scale, different substances can be meaningfully measured and compared as long as we choose the same scale. This stands in stark contrast to the characteristics of evolutionary independence highlighted by the heterogeneity objection.

I argue that the first three of these dissimilarities are not relevant here. The question is whether, as the heterogeneity objection implies, theoretical vagueness, heterogeneity of measurands, and conflicting operationalisations are good reasons to abandon the attempt to measure a particular theoretical quantity. To the extent that temperature was characterised by these three features, it is relevant to answering this question with respect to evolutionary independence. More interestingly, I will argue that the final dissimilarity – that temperature is independent from particular decisions of those who measure it – is false. While temperature is currently precise and independent from how scientists operationalise it, this was not always so. The remainder of this section will argue that, like evolutionary independence, temperature was vague, heterogeneous, and operationalised in multiple divergent ways.

To show that scientists measuring temperature originally faced the same difficulties as present-day taxonomists, consider first the vagueness of the theoretical notions of temperature and heat. Over the course of the seventeenth, eighteenth and part of the nineteenth century, during which much progress on the topic of heat was made, there was no generally accepted theoretical framework. Early

on, theories of heat were mostly based on Aristotle's thinking. They were over time replaced by caloric theories, which view heat as a substance. The distinction between heat and temperature only came to be recognised by Joseph Black in the second half of the eighteenth century, and it was not until well into the nineteenth century that caloric theories of heat were abandoned in favour of the kinetic theory that forms the basis of modern theories about heat and temperature (Barnett 1956; Chang 2004).

Second, the heterogeneity of measurands was arguably larger in the case of temperature than in that of evolutionary independence. According to Aristotle's thinking, which was still highly influential when the first thermometers were developed, hot and cold were even fundamental qualities that were present in different proportions in any body whatsoever (Barnett 1956). And just as in the case of evolutionary independence, heat is manifested in different substances in different ways and caused by a range of different processes. This heterogeneity is nicely illustrated by Francis Bacon's (1902, pp. 121–123) famous table of 'instances agreeing in the form of heat'. Besides unsurprising phenomena such as flames, hot springs and the rays of the sun, this table contains 'green and moist vegetable matter confined and rubbed together', 'the oil of marjoram', alcohol, vinegar when applied to wounded parts of the body, and even 'severe and intense cold'. Of course, we know now that all instances of heat are common in an interesting way. However, as this was not known to seventeenth and eighteenth century scientists developing thermometers, they faced heterogeneity at least as large as that observed today in evolutionary independence.

Finally, different scientists operationalised temperature in widely different ways, and results differed greatly depending on which measuring method was used (see Barnett 1956; and Chang 2004 for an overview). Over the course of more than two centuries, measurement procedures involved gas thermometers, fluid thermometers and hybrids between the two; a wide variety of thermometric substances including water, air, mercury, vinegar, alcohol and linseed oil; a variety of glass bulbs and containers, including Galileo's single bulb thermometer and later two-bulbed thermometers (Barnett 1956, p. 275); a variety of scales, such as Fahrenheit's, Celsius' and Newton's; and a variety of fixed points to calibrate these scales, including the boiling and freezing points of water, the melting point of butter, blood heat, first night frost, and the mean temperature of the King's chamber in the Great Pyramid of Giza (Chang 2004, p. 11).

Again, with our current insight in temperature it is tempting to think that all research using these various operationalisations was backed up by common principles. After all, many relied on the

expansion of thermometric substances when heated. We now know that temperature can be measured relatively reliably with a variety of thermometric substances, and that the mechanisms underlying these procedures are similar. In that sense, it might seem that there was more theoretical agreement about measuring temperature in the eighteenth century than there is now about assessing evolutionary independence. However, this seeming unity is an anachronism. Even if in hindsight their methods strike us as similar, scientists at the time were convinced that there was only one correct thermometric substance, and employed different candidates based on different theoretical assumptions. Without the current knowledge about the nature of temperature and the mechanisms underlying thermometers, scientists faced vagueness and heterogeneity similar to what currently characterises research on evolutionary independence.

In short, temperature was theoretically vague, heterogeneous, and operationalised in many ways. Still, this did not stop scientists from developing a highly precise and successful measure. This shows that even when the world turns out 'much messier than scientists would have liked' (Chang 2004, p. 49), sustained scientific effort can yield a precise and productive notion. This is why Hasok Chang's celebrated account of the history of temperature is called *Inventing temperature*: The precision of the concept is a scientific accomplishment and manufacturing it took multiple centuries of effort of many clever scientists. The point then is that heterogeneity, theoretical vagueness and multiple operationalisations need not stop us from developing a similarly useful measure of evolutionary independence.

Measurement as the way out

The previous section relied on the history of measuring temperature to argue that extensive heterogeneity does not make the measurement of a quantity impossible. This section continues to draw on the case of temperature to make the more positive point that measurement may even help to make the notion of evolutionary independence more precise.

The development of a precise theory and measure of temperature required overcoming the same catch-22 that taxonomists face today: to develop reliable measures, one needs a clear theory of what is measured; at the same time, such a theory can only be developed and tested if a reliable measure of that quantity is already at hand. This circular interdependence of theory and measurement is, of course, a well known problem in philosophy of science, which Van Fraassen (2008, chapter 5) and Tal (2013, pp. 1159–1160) have called the 'problem of coordination', and Chang (2004, pp. 57–60) the problem of 'nomic measurement'. Because of the impressive theoretical and metrological progress

over the course of only a few centuries, the case of temperature has received particular attention in discussions about the problem of coordination. While the details of this discussion exceed the scope of this paper, I want to highlight one important point of agreement that is particularly relevant here: developing measurement procedures, even in the absence of a clear theory, was crucial for overcoming the problem of coordination and developing modern theories of temperature. Similarly, I argue, pursuing measurement of evolutionary independence may be a promising strategy for making the notion more precise.

The history of thermometry is riddled with episodes that illustrate the importance of measurement in overcoming the coordination problem. Here, one brief example will serve to make the point.[5] The example concerns one of the most illustrious conceptual advances in the study of heat, namely, Joseph Black's notion of latent heat. According to Sherry (2011), it was the success of Black's quantitative concept of heat capacity that provided the first justification that temperature could be treated as a cardinal (rather than ordinal) quantity. More precisely, Black's theory assumed temperature to be such a quantity, and the success of his theory thus provided abductive support for this assumption. Important here is that Black was only successful at quantifying his notion of heat capacity because he could rely on Fahrenheit's very precise mercury thermometer to measure the change in temperature in mixtures of substances with different temperatures. Barnett (1956, p. 312) points out that experiments like Black's, which relied on very precise measurements, would not have been possible with the air thermometers that preceded the closed liquid thermometers such as Fahrenheit's. While Fahrenheit's thermometers suffered from the same theoretical problems as these air thermometers (i.e. they were not comparable to other measuring devices using different scales and different thermometric substances) they were far more precise and comparable to each other. Fahrenheit was known for his careful construction of thermometers, which would yield results that varied less than one sixteenth of a degree from each other (Sherry 2011, p. 512). This precision allowed Black to make the measurements required to quantify his notion of latent heat and develop his theory.

This episode illustrates that measurement in the absence of a developed theoretical framework can play a crucial role in resolving the coordination problem. The carefully crafted and highly comparable thermometers of Fahrenheit were a necessary requirement for Joseph Black's quantification of temperature and the concept of latent heat. Thus, at least in the case of temperature it would have

[5] For another apt example, see Chang's (2004, chapter 2) discussion of how Regnault showed the air thermometer to be more reliable than the mercury thermometer without making substantial theoretical commitments.

been wrong to assume that a developed theory was needed before engaging in measurement. Indeed, the development of reliable measurement procedures and a scale were necessary steps in getting to such a theory. Measurement theorists infer from this that the circularity between theory and measurement of the coordination problem is not vicious, but a way to overcome the problem: theory and operationalisation inform each other through multiple iterative steps. Rudimentary measurement procedures lead to outcomes that need to be accounted for theoretically; these theories, in turn, can then be used to construct better measurement methods, and so on. As van Fraassen (2008, p. 116) puts it: 'The questions *What counts as a measurement of (physical quantity) X?* and *What is (that physical quantity) X?* cannot be answered independently of each other'.

This suggests an alternative way forward for species delimitation: instead of abandoning measurement of evolutionary independence because of the messy state of nature, we can also pursue measurement to *create* a notion that is more precise. This may be possible, because measures and a scale like the Tobias criterion can be developed without a clear theory of evolutionary independence. These measures can then be one small step in the process of developing a clear theory. While there is no guarantee that this will turn out as successfully for evolutionary independence as it did for temperature, it is clear that the current vagueness and heterogeneity of the concept are not good reasons to think that it would not.

A plea for pragmatism

The previous section argued that measuring evolutionary independence may help to make the notion more precise in the same way measurement was crucial to clarifying temperature. I am not claiming, however, that measurement *will* turn out as successful for evolutionary independence as it did for temperature. Even if concepts like temperature are scientific creations, their success is also constrained by what the world is like and various historical contingencies. However, this need not discourage attempts to measure evolutionary independence. I argue that even if precision such as that of temperature turns out unattainable, sophisticated measures of evolutionary independence may be practically useful.

There are many useful scientific concepts that are vague, heterogeneous, and have a different meaning depending on the context of use and the way they are operationalised. Take the concept of inflation. Broadly defined as the increase in the general price level in an economy over a particular time, inflation is measured by tracking the change in a price index that in turn tracks the cost of buying a fixed basket of goods and services. However, depending on which goods are included in the basket

and how they are weighted, the concept takes on different meanings and measures yield different outcomes (Reiss 2008, 2013). As a consequence, it is generally accepted that 'there is no "true" inflation rate' but rather 'numerous inflation indices that are more or less useful relative to the given purpose' (Reiss 2008, p. 46). Still, the measurement of inflation plays a crucial role both in economic policymaking and in scientific research (Reiss 2008, pp. 23–24). Important here is that while the choice of any particular way of measuring inflation is partially arbitrary, economists and policymakers can usefully compare the price level at different points in time once one such a measure is chosen.

The analogy with inflation here serves to support a plea for pragmatism: a measure can be highly valuable even if it is recognised to be arbitrary. Arbitrariness would be a small price to pay if measuring evolutionary independence contributes to resolving taxonomic disorder. This plea ties into the debate about the use of DNA barcoding for species recognition and delimitation (Wheeler 2005; Hebert and Gregory 2005). Proponents of barcoding approaches emphasise epistemic values such as the gain in efficiency, comparability, and repeatability while opponents point out that barcoding fails to provide in-depth understanding of the complex evolutionary processes and relations that shape the organic world. A sophisticated measurement approach could yield precisely the epistemic advantages highlighted by proponents of DNA barcoding. Moreover, such an approach does not rule out the in-depth research advocated by barcoding opponents. Such research would be necessary to inform measurement approaches, particularly when they are attuned to particular taxa such as the Tobias criterion. In addition, thorough systematic research remains valuable even if there are constraints on which groups can be *recognised* as species.

Some taxonomists interpret the yardstick approach to evolutionary independence in this pragmatic way. Discussing 'pragmatic approaches' such as the Tobias criterion, Zachos (2018, p. 814) writes:

> Rather than continue to search for the Holy Grail, it has been suggested to agree on a consistent and quantifiable delimitation procedure for what is then called species. This way, given the same raw data, different taxonomists would at least come up with the same species delimitations in a consistent and repeatable (albeit still not completely nonarbitrary) way.

In other words, a practically useful measure of evolutionary independence, along with a threshold for species status, may be the best we can do. Such a measure could be attuned and limited to certain taxa, such as mammals, to ensure that at least within these groups species delimitation would be transparent, intersubjectively stable, and yield units that are comparable for the properties that are chosen for the standardised operationalisation. While these properties may not be more interesting

than the ones used by other viable operationalisations, this would at least guarantee that research or practices that rely on species as their currency are not working with radically different groups of organisms. This is precisely the approach that economists have taken, with considerable practical success, to measuring inflation.

Note that the practical usefulness of such a measure for evolutionary independence should be evaluated in comparison with current practices of species recognition. That is, a measurement system does not have to resolve all issues of the species problem to be worthwhile; instead, it must simply be better, all things considered, than the current system and its disorder. This is precisely what Tobias et al. (2010, p. 742) state in support of their measurement method:

> [The Tobias Criterion] clearly adds a greater measure of uniformity to the taxonomic decision making process, and has the power to produce taxonomic changes that are consistent and easily valuated by independent reviewers. This contrasts markedly with current practice in avian systematics, which generates anything from narrowly divergent allopatric 'species' to highly divergent 'subspecies'. If carefully applied, our system can therefore help to resolve difficult cases with conservation implications, and to produce a global taxonomy of comparable species units.

Many scientific concepts, such as inflation, well-being and perhaps even ecological fitness, show that pursuing measurement of vague concepts can yield considerable epistemic and practical benefits. Again, the point is that the heterogeneity and vagueness of evolutionary independence do not provide good reasons to abandon the measurement approach. On the contrary, even: measurement methods can sometimes create an acceptable degree of comparability where it is lacking in the world.

Conclusions

I have argued that the vagueness of the concept of evolutionary independence and the heterogeneity of phenomena it applies to do not provide good reasons to abandon attempts to measure evolutionary independence. Indeed, measurement may even help in further clarifying the notion, and might yield epistemic benefits even if precision such as that of temperature is unattainable. To end the paper, I want to briefly point out that the measurement approach fits well with recent developments in the field of species delimitation.

This field has recently witnessed a remarkable explosion in new and sophisticated methods for delimiting species (Sites and Marshall 2003; Camargo and Sites 2013; Carstens et al. 2013). These

methods mostly use genetic and genomic data in combination with evolutionary models to pick out independently evolving lineages. Characteristic of this 'renaissance' (Sites and Marshall 2003, p. 462) in species delimitation is its focus on the operationalisation of species classification. Rather than engaging in controversies about species concepts, these new approaches in taxonomy bracket theoretical debates, and instead focus on practical methods of delimitation. Even though these methods often yield conflicting results and thus continue taxonomic disorder, the results are promising: the models become increasingly sophisticated and are able to take into account more relevant variables, the procedures for any single method are transparent and yield the same results regardless of who performs them, and these methods improve our understanding of speciation and the various processes that play a role in it (Fujita et al. 2012; Flot 2015).

The measurement approach fits in this new tradition in that it tries to resolve taxonomic disorder by focusing on operationalisation rather than species concepts. This is not to imply that we should not try to make the theoretical notion of evolutionary independence more precise.[6] Instead, the point is that the easiest way of doing this may be through developing measurement procedures. With its focus on operationalisation over theory, the position defended in this paper (and taken from the recent renaissance) is the opposite of that defended recently about measurement of mental attributes (Bringmann and Eronen 2016) and well-being (Alexandrova and Haybron 2016). These authors argue that in both fields a lack of theoretical development combined with nearly exclusive attention for operationalisations impedes progress in tracking these properties. With the existing sophisticated measures in hand, they argue, it has become clear that theory-avoidant focus on operationalisations is not productive. I argue that species delimitation is on the other side of the cycle between theory and measurement: pursuing sophisticated measurement methods may help us break through the theoretical deadlock of the species problem and provide a solution for taxonomic disorder.

References

Alexandrova A, Haybron D (2016) Is Construct Validation Valid? Philosophy of Science 83:1089–1109

Balen S, Eaton JA, Rheindt FE (2013) Biology, taxonomy and conservation status of the Short-tailed Green Magpie Cissa [t.] thalassina from Java. Bird Conservation International 23:91–109. https://doi.org/10.1017/S0959270911000360

[6] For two interesting recent attempts at going beyond metaphorical characterisations of evolutionary lineages, see Barker (2019) and Sterner (2017, 2019). Notably, both Sterner and Barker pay explicit attention to how their accounts can be operationalised.

Barker MJ (2019) Species and Other Evolving Lineages as Feedback Systems. Philosophy, Theory, and Practice in Biology 11:. http://dx.doi.org/10.3998/ptpbio.16039257.0011.013

Barker MJ (2007) The empirical inadequacy of species cohesion by gene flow. Philosophy of Science 74:654–665. https://doi.org/10.1086/525611

Barnett MK (1956) The Development of Thermometry and the Temperature Concept. Osiris 12:269–341. https://doi.org/10.1086/368601

Baum DA (2009) Species as ranked taxa. Syst Biol 1:74–86. https://doi.org/10.1093/sysbio/syp011

Bringmann LF, Eronen MI (2016) Heating up the measurement debate: What psychologists can learn from the history of physics. Theory & Psychology 26:27–43. https://doi.org/10.1177/0959354315617253

Burfield IJ, Butchart SHM, Collar NJ (2017) BirdLife, conservation and taxonomy. Bird Conservation International 27:1–5. https://doi.org/10.1017/S0959270917000065

Camargo A, Sites J (2013) Species delimitation: A decade after the renaissance. In: Pavlinov I (ed) The Species Problem - Ongoing Issues. InTech, Rijeka, pp 225–247

Carstens BC, Pelletier TA, Reid NM, Satler JD (2013) How to fail at species delimitation. Mol Ecol 22:4369–4383. https://doi.org/10.1111/mec.12413

Cartwright N, Bradburn NM, Fuller J (2011) A theory of measurement. Centre for Humanities Engaging Science and Society (CHESS), Durham

Chang H (2004) Inventing Temperature: Measurement and Scientific Progress. Oxford University Press, New York

Conix S (2018) Integrative taxonomy and the operationalization of evolutionary independence. European Journal for Philosophy of Science 8:587–603. https://doi.org/10.1007/s13194-018-0202-z

Conix S (2019) In defence of taxonomic governance. Org Divers Evol. https://doi.org/10.1007/s13127-019-00391-6

De Queiroz K (1998) The general lineage concept of species, species criteria, and the process of speciation: a conceptual unification and terminological recommendations. In: Howard D, Berlocher S (eds) Endless Forms: Species and Speciation. Oxford University Press, Oxford, UK, pp 57–75

Donegan TM (2018) What is a species? A new universal method to measure differentiation and assess the taxonomic rank of allopatric populations, using continuous variables. ZooKeys 757:1–67. https://doi.org/10.3897/zookeys.757.10965

Ehrlich PR, Raven PH (1969) Differentiation of populations. Science 165:1228–1232. https://doi.org/10.1126/science.165.3899.1228

Faurby S, Eiserhardt WL, Svenning J-C (2016) Strong effects of variation in taxonomic opinion on diversification analyses. Methods Ecol Evol 7:4–13. https://doi.org/10.1111/2041-210X.12449

Flot J-F (2015) Species Delimitation's Coming of Age. Syst Biol 64:897–899. https://doi.org/10.1093/sysbio/syv071

Fraassen BCV (2008) Scientific Representation: Paradoxes of Perspective. Oxford University Press, Oxford

Fujita MK, Leaché AD, Burbrink FT, et al (2012) Coalescent-based species delimitation in an integrative taxonomy. Trends in Ecology & Evolution 27:480–488. https://doi.org/10.1016/j.tree.2012.04.012

Galtier N (2019) Delineating species in the speciation continuum: a proposal. Evolutionary Applications 0: https://doi.org/10.1111/eva.12748

García-París M, Jockusch EL (1999) A mitochondrial DNA perspective on the evolution of Iberian Discoglossus (Amphibia: Anura). Journal of Zoology 248:209–218

Garnett ST, Christidis L (2017) Taxonomy anarchy hampers conservation. Nature 546:25–27. https://doi.org/10.1038/546025a

Godfray HCJ (2002) Challenges for taxonomy. Nature 417:17–19. https://doi.org/10.1038/417017a

Haber MH (2019) Species in the Age of Discordance. Philosophy, Theory, and Practice in Biology 11:. http://dx.doi.org/10.3998/ptpbio.16039257.0011.021

Haber MH (2012) Multilevel Lineages and Multidimensional Trees: The Levels of Lineage and Phylogeny Reconstruction. Philosophy of Science 79:609–623. https://doi.org/10.1086/667849

Halley MR, Klicka JC, Clee PRS, Weckstein JD (2017) Restoring the species status of Catharus maculatus (Aves: Turdidae), a secretive Andean thrush, with a critique of the yardstick approach to species delimitation. Zootaxa 4276:387–404. https://doi.org/10.11646/zootaxa.4276.3.4

Hebert P, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Systematic Biology 54:852–859. https://doi.org/10.1080/10635150500354886

Hey J, Pinho C (2012) Population genetics and objectivity in species diagnosis. Evolution 5:1413–1429

Hoyo J del, Collar NJ (2014) HBW and BirdLife International illustrated checklist of the birds of the world. Lynx Edicions, Barcelona

Isaac NJB, Mallet J, Mace GM (2004) Taxonomic inflation: Its influence on macroecology and conservation. Trends in Ecology & Evolution 19:464–469. https://doi.org/10.1016/j.tree.2004.06.004

Khuroo AA, Dar GH, Khan ZS, Malik AH (2007) Exploring an inherent interface between taxonomy and biodiversity: Current problems and future challenges. Journal for Nature Conservation 15:256–261. https://doi.org/10.1016/j.jnc.2007.07.003

Mace GM (2004) The role of taxonomy in species conservation. Phil Trans R Soc Lond B 359:711–719. https://doi.org/10.1098/rstb.2003.1454

Maddison WP (1997) Gene Trees in Species Trees. Syst Biol 46:523–536. https://doi.org/10.1093/sysbio/46.3.523

Mallet J, Besansky N, Hahn MW (2015) How reticulated are species? BioEssays 140–149. https://doi.org/10.1002/bies.201500149

Mari L, Carbone P, Giordani A, Petri D (2017) A structural interpretation of measurement and some related epistemological issues. Studies in History and Philosophy of Science Part A 65–66:46–56. https://doi.org/10.1016/j.shpsa.2017.08.001

Mayden R (1997) A hierarchy of species concepts: the denouement in the saga of the species problem. In: Claridge M, Dawah H, Wilson RA (eds) Species, the Units of Biodiversity, Systematics Association Special Volume Series. Chapman & Hall, London, pp 381–424

Remsen JV (2016) A “rapid assessment program” for assigning species rank? Journal of Field Ornithology 87:110–115. https://doi.org/10.1111/jofo.12142

Remsen JV (2015) Book Review: HBW and BirdLife International Illustrated Checklist of the Birds of the World Volume 1: Non-passerines. Journal of Field Ornithology 86:182–187. https://doi.org/10.1111/jofo.12102

Rheindt FE, Székely T, Edwards SV, et al (2011) Conflict between Genetic and Phenotypic Differentiation: The Evolutionary History of a ‘Lost and Rediscovered’ Shorebird. PLOS ONE 6:e26995. https://doi.org/10.1371/journal.pone.0026995

Schlick-Steiner BC, Steiner FM, Seifert B, et al (2010) Integrative taxonomy: A multisource approach to exploring biodiversity. Annual Review of Entomology 55:421–438. https://doi.org/10.1146/annurev-ento-112408-085432

Schuchmann K-L, Weller A-A, Jürgens D (2016) Biogeography and taxonomy of racket-tail hummingbirds (Aves: Trochilidae: Ocreatus): evidence for species delimitation from morphology and display behavior. Zootaxa 4200:83–108. https://doi.org/10.11646/zootaxa.4200.1.3

Sherry D (2011) Thermoscopes, thermometers, and the foundations of measurement. Studies in History and Philosophy of Science Part A 42:509–524. https://doi.org/10.1016/j.shpsa.2011.07.001

Simpson G (1961) Principles of Animal Taxonomy. Columbia University Press, New York, NY

Sites JW, Marshall JC (2003) Delimiting species: a Renaissance issue in systematic biology. Trends in Ecology & Evolution 18:462–470. https://doi.org/10.1016/S0169-5347(03)00184-8

Sterner B (2017) Individuating population lineages: a new genealogical criterion. Biol Philos 32:683–703. https://doi.org/10.1007/s10539-017-9580-4

Sterner BW (2019) Evolutionary Species in Light of Population Genomics. Philosophy of Science. https://doi.org/10.1086/705527

Tal E (2013) Old and New Problems in Philosophy of Measurement. Philosophy Compass 8:1159–1173. https://doi.org/10.1111/phc3.12089

Templeton (1992) The meaning of species and speciation: a genetic perspective. In: Ereshefsky M (ed) The Units of Evolution: Essays on the Nature of Species. MIT Press, Cambridge, MA, pp 159–183

Tobias JA, Seddon N, Spottiswoode CN, et al (2010) Quantitative criteria for species delimitation. Ibis 152:724–746. https://doi.org/10.1111/j.1474-919X.2010.01051.x

Vallejo-Marín M, Buggs RJA, Cooley AM, Puzey JR (2015) Speciation by genome duplication: Repeated origins and genomic composition of the recently formed allopolyploid species Mimulus peregrinus. Evolution 69:1487–1500. https://doi.org/10.1111/evo.12678

Wheeler QD (2005) Losing the plot: DNA “barcodes” and taxonomy. Cladistics 21:405–407. https://doi.org/10.1111/j.1096-0031.2005.00075.x

Wiley EO (1978) The evolutionary species concept reconsidered. Syst Biol 27:17–26. https://doi.org/10.2307/2412809

Willis SC (2017) One species or four? Yes!...and, no. Or, arbitrary assignment of lineages to species obscures the diversification processes of Neotropical fishes. PLOS ONE 12:e0172349. https://doi.org/10.1371/journal.pone.0172349

Willmann R, Meier R (2000) A critique from the hennigian species concept perspective. In: Wheeler QD, Meier R (eds) Species Concepts and Phylogenetic Theory: A Debate. Columbia University Press, New York, NY, pp 101–119

Zachos FE (2016) Species Concepts in Biology: Historical Development, Theoretical Foundations and Practical Relevance. Springer, Basel

Zachos FE (2018) (New) Species concepts, species delimitation and the inherent limitations of taxonomy. J Genet 97:811–815. https://doi.org/10.1007/s12041-018-0965-1