IBIS The International Journal of Avian Science

Ibis (2010), **152**, 724–746

Quantitative criteria for species delimitation

JOSEPH A. TOBIAS,[1]* NATHALIE SEDDON,[1] CLAIRE N. SPOTTISWOODE,[2] JOHN D. PILGRIM,[3] LINCOLN D. C. FISHPOOL[3] & NIGEL J. COLLAR[3]

[1]*Department of Zoology, Edward Grey Institute, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*

[2]*Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK*

[3]*BirdLife International, Wellbrook Court, Girton Road, Cambridge CB3 0NA, UK*

Species are the fundamental units of biology, ecology and conservation, and progress in these fields is therefore hampered by widespread taxonomic bias and uncertainty. Numerous operational techniques based on molecular or phenotypic data have been designed to overcome this problem, yet existing procedures remain subjective or inconsistent, particularly when applying the biological species concept. We address this issue by developing quantitative methods for a classic technique in systematic zoology, namely the use of divergence between undisputed sympatric species as a yardstick for assessing the taxonomic status of allopatric forms. We calculated mean levels of differentiation in multiple phenotypic characters – including biometrics, plumage and voice – for 58 sympatric or parapatric species-pairs from 29 avian families. We then used estimates of mean divergence to develop criteria for species delimitation based on data-driven thresholds. Preliminary tests show that these criteria result in relatively few changes to avian taxonomy in Europe, yet are capable of extensive reassignment of species limits in poorly known tropical regions. While we recognize that species limits are in many cases inherently arbitrary, we argue that our system can be applied to the global avifauna to deliver taxonomic decisions with a high level of objectivity, consistency and transparency.

Keywords: avian taxonomy, conservation units, operational guidelines, phenotypic divergence, species limits.

Biodiversity can be subdivided into a range of categories from genes to ecosystems, but it is the species category in particular that underpins much of biology, ecology and conservation (May 1990, Mace 2004, de Queiroz 2005). Although it can be argued that species limits are not important for understanding evolution, or even speciation (Winker *et al.* 2007, Mallet 2008), they have a major influence on biogeography, community ecology and macroecology, all of which rely on estimates of species richness, abundance and distribution (Agapow *et al.* 2004, Isaac *et al.* 2004, Tobias *et al.* 2008). Moreover, species are crucial to conservationists and policy-makers, who use them as units for prioritizing action and formulating national and international law, and who therefore require species delimitation to be consistent and transparent (Ryder 1986, Collar 1997, Agapow 2005, Garnett & Christidis 2007).

The emphasis on species is here to stay for a variety of biological and cultural reasons, and yet the same can be said of the 'species problem' (Hey 2001, Hey *et al.* 2003). Biodiversity tends to segregate neatly into discrete species at any given locality, and therefore taxonomic confusion is negligible in sympatry. However, the wider picture is greatly complicated by geographical variation and the gradual evolution of reproductive isolation in allopatry. Lineages diverge slowly, and taxonomy involves placing cut-offs somewhere along the transition from populations to species. For this reason, amongst others, it is generally agreed that no species concept can fully capture the arbitrary nature of species boundaries (Stebbins 1969, Hendry *et al.* 2000, Hey 2006, Wiens 2007, Joseph & Omland 2009). This element of subjectivity means that species classification is best viewed as a

*Corresponding author.
Email: joseph.tobias@zoo.ox.ac.uk

'man-made system of pigeonholes' devised for the purpose of subdividing biotic variation into convenient units (Dobzhansky 1937). To ensure that these 'pigeonholes' are comparable entities, we need a man-made system of taxonomic criteria that maximizes consistency and transparency, and hence objectivity.

Decisions in avian taxonomy were once relatively uncontested matters for a small museum-based community, but in the past few decades several factors have tended to destabilize species lists. For example, the biological species concept (BSC), whose cornerstone principle is reproductive isolation (Dobzhansky 1937, Mayr 1942), has been partly replaced by various alternatives. These include the phylogenetic species concept (PSC), whose cornerstone is diagnosability (Cracraft 1989), and the monophyletic species concept (MSC), whose cornerstone is monophyly (Mishler & Donoghue 1982). The instability caused by shifting species concepts has been accelerated by the rapid development of molecular analysis, which has led to major advances in our understanding of the historical descent of lineages, but also to widespread acknowledgement that degrees of genetic divergence cannot easily be translated into species limits (Edwards *et al.* 2005, Winker *et al.* 2007, Joseph & Omland 2009, Winker 2009). Meanwhile, other sources of bias and confusion have emerged simply because there is broader participation in the taxonomic evaluation process, involving academics, national committees, amateur naturalists, conservationists, and the authors of textbooks and field guides (e.g. Rowlett 2003, Chaitra *et al.* 2004).

The main outcome of these changes is taxonomic uncertainty, which in turn has led to a proliferation of quantitative methods designed to improve the rigour of species delimitation. Many of these techniques are of limited utility, however, because they are geared to comparing taxa on the basis of detailed phenotypic (Wiens & Servedio 2000) and genetic datasets (Sites & Marshall 2003, 2004, Knowles & Carstens 2007, Shaffer & Thomson 2007). To be effective, these datasets need to be sampled from multiple populations, individuals and loci, and they are therefore often expensive and time-consuming to compile. Proponents of genetic barcoding have sought to accelerate the process by delimiting species en masse (Hebert *et al.* 2004), but barcodes cannot replace taxonomy for a variety of reasons, not least because they reflect a single locus or character (see Moritz & Cicero 2004, DeSalle *et al.* 2005, Meyer & Paulay 2005, Will *et al.* 2005, Elias *et al.* 2008). Meanwhile, we face escalating threats to the environment and continuing uncertainty about how many species it supports (May 1988, Collar 2003, Tobias *et al.* 2008).

One possible solution is to refine the classic technique of using the degree of phenotypic divergence in sympatric species to guide judgements about allopatric taxa. The logic of comparing divergence in putative species against that found between undisputed sympatric species can be traced back to leading European ornithologists of the early 20th century, including Ernst Hartert and Walter Rothschild. The procedure was later championed by Ernst Mayr (1969). In its traditional format, it involved simple visual comparisons of museum specimens, and was therefore open to subjectivity and misinterpretation. It was not until the end of the 20th century that it was rephrased in a quantitative framework. This step was taken by Isler *et al.* (1998), who established criteria for species designation in antbirds (Thamnophilidae) based on song divergence in eight sympatric species-pairs. This approach laid the foundations for a series of taxonomic re-appraisals (Isler *et al.* 1999, 2005, 2007, 2009, Braun *et al.* 2005), but it is only relevant to antbirds and antbird vocalizations (Zink 2006).

An attempt to introduce broader procedural consistency was made by the British Ornithologists' Union in its 'guidelines for the application of species limits to sympatric, parapatric, allopatric and hybridizing taxa' (Helbig *et al.* 2002, hereafter 'the BOU guidelines'). These proposals – summarized in the Supporting Information Table S1 – were based on quantifying the number of differences between pairs of taxa, and making a rough comparison of their phenotypic and genetic divergence with that found in related pairs of sympatric species. The BOU guidelines succeeded in adding a quantitative dimension to Mayr's (1969) procedure, and a degree of uniformity and efficiency to the taxonomic decision-making process. However, the guidelines lack clarity in key areas, including the methods for divergence evaluation. Most importantly, they provided no explicit recommendations for (i) judging the number or magnitude of differences required to trigger species status, (ii) limiting the influence of trivial differences, or (iii) comparing divergence in different types of trait, e.g. songs vs. plumage. A more detailed

appraisal of the BOU guidelines is provided in Appendix S1.

In this paper, we build on the advances of Isler *et al.* (1998) and the BOU guidelines by developing a revised set of criteria based on direct quantification of divergence in multiple traits. We assess divergence in a sample of sympatric species, then use this to calibrate thresholds for species status in allopatric or parapatric taxa. Finally, we discuss the strengths and weaknesses of the system based on the results of empirical tests against current subspecies limits. The main goal is to produce a system that is simple and rapid enough to be applied to all birds, quantitative enough to satisfy the need for rigour and repeatability, and transparent enough to allow taxonomic decisions to be traced and evaluated. Before presenting our system of thresholds we outline our position regarding several relevant issues, including species concepts, paraphyly and the use of genetic data.

SPECIES CONCEPT

The PSC, MSC and BSC all fall under the general lineage concept of species (de Queiroz 1998). They are not so much species conceptualizations as criteria for recognizing species, each emphasizing a different line of evidence for lineage separation (de Queiroz 2007). Discrepancies between them arise because they focus on different stages of the speciation process. A typical sequence of events for two recently separated lineages begins with the evolution of diagnostic differences (the hallmark of the PSC), then reciprocal monophyly (the hallmark of the MSC) and finally reproductive isolation (the hallmark of the BSC). Thus, any system of taxonomic guidelines must begin by defining which of these standards will be used to delimit species.

The PSC and MSC are valuable tools for studying evolution and diversification, but they have some drawbacks as a basis for species taxonomy. First, the criterion of reciprocal monophyly is not always easy to pin down because each gene has its own evolutionary history that is not necessarily congruent with that of other genes, or indeed populations (Coyne & Orr 2004). Secondly, when using a rapidly evolving gene like mitochondrial DNA (mtDNA), monophyly can either arise very early in the speciation process or long after speciation is complete, depending on whether populations have the opportunity to hybridize (Funk & Omland 2003, Irwin *et al.* 2009, Joseph & Omland 2009). Thirdly, strong inferences of monophyly are only possible when geographically intermediate populations have been adequately sampled (Remsen 2005). Fourthly, the criterion of diagnosability is problematical because there is no clear limit to how subtle a diagnostic difference can be, which opens the door to unconstrained taxonomic inflation via character triviality (Collar 1997, Johnson *et al.* 1999, Isaac *et al.* 2004, Mace 2004, Garnett & Christidis 2007, Winker *et al.* 2007). This is a serious consideration in tropical archipelagos or mountain ranges, where a vast number – many tens of thousands – of subtly divergent, sedentary populations are likely to be monophyletic (Phillimore & Owens 2006, Phillimore *et al.* 2008), or diagnosable by at least one minor trait (Collar 1997, Price 2008).

Our criteria are based on the BSC, not because it is 'right' but because it has some advantages as a framework for global taxonomic treatments. In particular, it relies on the semi-objective criterion of reproductive isolation, which applies a relatively fixed and broadly intuitive limit to species diversity. These are important considerations, particularly for conservationists and legislators (Collar 1997, Mace 2004, Winker *et al.* 2007). The BSC is often challenged on the grounds that it struggles to deal with hybridization, and distorts evolutionary history by grouping non-monophyletic populations together as taxa (Rosen 1978, Cracraft 1983, Donoghue 1985). These objections are justified but not fatal, first because modern applications of the BSC allow hybridization between species (e.g. Johnson *et al.* 1999, Helbig *et al.* 2002), and secondly because non-monophyletic species are often an accurate reflection of biological reality (see below). A more serious drawback is that decisions on the status of allopatric taxa are essentially subjective and arbitrary because without geographical contact there can be no direct test of reproductive isolation (Mayr 1942, Brown & Wilson 1956, Cracraft 1989, Zink & McKitrick 1995). Again, this need not be fatal as arbitrariness can be minimized using direct comparisons with sympatric species (Isler *et al.* 1998, Johnson *et al.* 1999, Winker 2010, this study).

PARAPHYLY AND HYBRIDIZATION

Paraphyly, which is anathema to many cladistic systematists, lies at the core of many taxonomic

disputes. Paraphyly in gene trees often reflects (i) incomplete lineage sorting (i.e. the failure of gene lineages to fix along species lineages), or (ii) hybridization and selective introgression (Funk & Omland 2003, Wang *et al.* 2008, Joseph & Omland 2009, McKay & Zink 2010). These are common features of biological systems because hybridization may continue long after speciation, and the sorting of ancestral polymorphisms may continue long after hybridization has ceased (Crandall *et al.* 2000, Edwards *et al.* 2005, Rosenberg 2007, Carling & Brumfield 2008, Wang *et al.* 2008). Thus, paraphyletic gene trees are widespread even in geographically or reproductively isolated lineages.

Perhaps more importantly, intraspecific paraphyly can be reflected not only in individual gene trees, but at the level of populations. Indeed, population-level paraphyly appears to be commonplace in the case of peripatric or 'budding' speciation (Frey 1993), in which reproductive isolation evolves rapidly in a spatially isolated population embedded within a widespread taxon, leading to paraphyly of the ancestor (Harrison 1998, see fig. 2 in Funk & Omland 2003). This may occur, for example, when a continental form colonizes an island, or expands its range into different habitats, such that ecological selection drives phenotypic evolution, and thereby speciation (see Losos *et al.* 1997, Schluter 2009). Evidence of this form of paraphyly can be found in a wide variety of taxa, including birds (Talbot & Shields 1996, Hedin 1997, Omland *et al.* 2000, Salzburger *et al.* 2002, Zink *et al.* 2009).

There are two schools of thought regarding the taxonomic treatment of population-level paraphyly. One is that it reflects 'incorrect taxonomy' and can be rectified simply by matching species limits to monophyletic groupings based on mtDNA (McKay & Zink 2010). Many recent papers propose splits on this basis. For example, it has been suggested that *Corvus corax clarionensis* should be elevated to species level to avoid paraphyly with respect to *Corvus cryptoleucus* (McKay & Zink 2010), or that the *Motacilla flava/Motacilla citreola* complex requires further subdivision to eliminate paraphyly (Pavlova *et al.* 2003), despite the fact that these revisions would break up phenotypically homogeneous groupings.

We take the opposing view, that a distinctive, reproductively isolated lineage can be classified as a species even though it is nested within a phenotypically homogeneous ancestor. To clarify, if subspecies A and B are phenotypically similar, but genetically and geographically interposed by a third divergent and reproductively isolated taxon C, it does not follow that the classification of C as a separate species must necessarily trigger the splitting of A and B (Lee 2003, Coyne & Orr 2004, Nordal & Stedje 2005, Zander 2007, Joseph & Omland 2009). It is clear that lumping non-sisters in this way results in a mismatch between species and clades. However, we concur with Lee (2003), who argued that 'this mismatch is precisely what makes the species category worthy of special recognition: species are *not* merely another type of clade, but a different type of biological entity altogether'. From this perspective, useful information is lost when taxonomy is forced to reflect gene trees by either over-lumping daughter and parent species, or over-splitting inherently paraphyletic taxa, and thereby ignoring the evolutionary reality of the nested lineage (see Hedin 1997, Harrison 1998, Funk & Omland 2003, Coyne & Orr 2004, Joseph & Omland 2009).

A related area of disagreement involves hybridization, which occurs between a surprisingly large proportion of avian species (Grant & Grant 1992, McCarthy 2006), even long after speciation (Price & Bouvier 2002, Mallet 2005). While it is therefore important that any concept of avian species admits a degree of hybridization (Johnson *et al.* 1999), this leaves open the question of where we should place the cut-off point between species and subspecies. No resolution to this issue is in sight, as demonstrated by numerous inconsistencies in the taxonomic treatment of hybridization: many distinctive taxa that hybridize frequently across broad contact zones are classified as species (e.g. *Melanerpes aurifrons* and *Melanerpes carolinus*; *Emberiza leucocephalos* and *Emberiza citrinella*; *Vermivora pinus* and *Vermivora chrysopterus*), whereas others with similar or reduced levels of hybridization are lumped (e.g. *Dendroica coronata coronata* and *Dendroica coronata auduboni*; *Pheugopedius nigricapillus nigricapillus* and *Pheugopedius nigricapillus castaneus*) (Price 2008, Brelsford & Irwin 2009, Irwin *et al.* 2009). The aim of our system is to classify hybridizing taxa as species where the degree of divergence between pure phenotypes exceeds a threshold set by known parapatric and sympatric species.

MOLECULAR TAXONOMY

Molecular data can reveal the historical descent of lineages and the extent of gene flow between them. These insights are relatively easy to interpret in the case of genera and families, and phylogenetic analyses are therefore revolutionizing higher-level systematics in birds (e.g. Chesser *et al.* 2007, Hackett *et al.* 2008). Genes are also informative in species-level taxonomy, particularly when populations meet, in which case divergence or monophyly in genetic markers indicates that mating is assortative, or that hybrids are inviable. In most cases, phylogenies are therefore capable of revealing cryptic species in geographical contact (e.g. Sorenson *et al.* 2003, Toews & Irwin 2008, Benkman *et al.* 2009, Brambilla *et al.* 2009), or near contact in the case of dispersive taxa (e.g. Cibois *et al.* 2007, Dávalos & Porzecanski 2009). Even so, it is worth bearing in mind that the gene trees of sympatric or parapatric populations regularly fail to reflect species limits as a result of incomplete lineage sorting, or hybridization and selective introgression (Ballard & Rand 2005, Alström *et al.* 2008a, Irwin *et al.* 2009).

Allopatric populations are also routinely delimited on the basis of relatively simplistic molecular approaches, such as mtDNA monophyly or 'diagnosability' (e.g. Zink 1994, García-Moreno & Fjeldså 1999, Abbott & Double 2003, Dietzen *et al.* 2008, Techow *et al.* 2009). This approach is much more contentious because, without geographical contact between populations, molecular data are less informative about species limits (Helbig *et al.* 1995, Sangster 2000a, Edwards *et al.* 2005, Wiens 2007, Winker 2010). Most importantly, mtDNA often sorts completely in isolated lineages long before the evolution of reproductive isolation or even phenotypic divergence (Helbig *et al.* 1995, Hendry *et al.* 2000, Phillimore *et al.* 2008, Price 2008). In other words, lineages are bound to qualify as independently evolving units even after a relatively brief allopatric phase. Treating all such entities as species can lead to taxonomic chaos, particularly in the montane or insular tropics where barriers to gene flow abound. In effect, a taxonomy based on this approach may reflect either (i) barriers to dispersal and gaps in distribution or (ii) biases in the availability of genetic data, rather than more fundamental attributes such as reproductive isolation or phenotypic divergence. Of course, many intraspecific lineages will become reproductively isolated in future as divergence continues, and thus our species limits should not be interpreted to mean that anything falling below these limits is unimportant to conservation. We argue that all independent evolutionary lineages should be treated as units of conservation significance, but not necessarily as species (Meiri & Mace 2007).

We do not mean to imply that genetic analyses cannot contribute to species delimitation of allopatric forms under the BSC. On the contrary, comparing molecular divergence with that found between irrefutable species is clearly useful inasmuch as it 'gives a rough indication of how likely it is that reproductive incompatibilities have evolved between two taxa' (Helbig *et al.* 2002). We therefore agree that taxonomic decisions should be based on the maximum number of available characters (de Queiroz 1998, 2007, Sites & Marshall 2004, Winker 2009), and that the ideal scenario involves a combination of ecological, behavioural, phenotypic and genotypic data (e.g. Yoder *et al.* 2005, Alström *et al.* 2008b, Leaché *et al.* 2009, Cadena & Cuervo 2010). However, molecular data are typically unavailable in the required format (i.e. sampled throughout species' ranges from multiple individuals and at multiple loci). Indeed, for most taxa, appropriate genetic samples may not be generated for decades, and thus a standardized molecular taxonomy is not yet feasible, even for a relatively well-known group like birds. In the meantime, the best available datasets for global treatments involve phenotypic characters.

In summary, molecular data are a mixed blessing for the BSC. On the one hand, we suggest that genes should take precedence as a taxonomic character in cases of known geographical contact between breeding populations, particularly when they reveal assortative mating. On the other hand, our system of criteria may provide a more useful framework for taxonomic decisions in cases of allopatry or extensive hybridization. In the version presented here, we focus on phenotypic divergence and exclude molecular divergence, largely because of the patchiness of genetic data and the extent of disagreement about how they should be applied to species limits (Edwards *et al.* 2005, Knowles & Carstens 2007, Price 2008, Joseph & Omland 2009, Winker 2009). Nonetheless, our system is designed to use any form of quantitative evidence, including thresholds of molecular divergence. We

hope that such thresholds can be developed and incorporated when more extensive comparative data are available (see Discussion).

METHODS

Our approach is based on measuring phenotypic divergence in undisputed species to establish thresholds for a taxonomic scoring system. The following sections serve as a rationale for this technique and as guidelines for its application; the background and methods we provide for quantifying divergence between sympatric and parapatric species can be directly transferred to comparisons between any pair of allopatric taxa. Note that thresholds set in the following section are based on data presented in the Results.

Throughout, we use 'trait' to mean a broad class of phenotypic differences (plumage, song, biometrics, etc.), and 'character' to mean any diagnostic difference identified by a cross-taxon comparison. We assume that characters (i) can be described, measured or counted and (ii) are consistently present in members of the same age and/or sex class of a population, on the basis of which that population can be classified as a taxon. The most relevant characters for our purposes are morphological (e.g. size and shape), visual (e.g. plumage colour and pattern), acoustic (e.g. pitch and pace) and behavioural (e.g. courtship display or nest type). It is worth observing here that allopatric disjunction, however great, is not in itself a character, and should not be taken as evidence of species status.

Species delimitation by phenotype

The use of phenotype to delimit species has some drawbacks but many advantages. In particular, phenotypic characters are often determined by multiple genes, and thus an assessment of numerous phenotypic characters is more likely to reflect divergence across the genome than are molecular methods dependent on limited sampling of loci. Of course, museum taxonomists have long been making use of phenotypic datasets, although they have typically been restricted to a partial set of characters visible in specimens. The BOU guidelines continued to emphasize the importance of divergence in morphological and plumage traits, but gave less space to other forms of phenotypic divergence, such as vocal and behavioural traits. In recent years, the availability of these additional datasets has improved greatly, and we believe that this paves the way for a new set of criteria that incorporates a more complete set of phenotypic characters in a quantitative framework.

The use of vocal characters in avian taxonomy has increased dramatically in recent decades for three main reasons (Isler *et al.* 1998, Alström & Ranft 2003). First, acoustic signals often play a central role in species recognition and mate choice, and therefore mediate reproductive isolation in many avian systems (Catchpole & Slater 1995, Baker & Boylan 1999, Slabbekoorn & Smith 2002, Marler & Slabbekoorn 2004, Price 2008, Toews & Irwin 2008). Indeed, vocal characters often provide a better indication of species limits and relationships than morphological characters (e.g. Martens *et al.* 2003, Päckert *et al.* 2003, Rheindt *et al.* 2008). Secondly, the use of songs and calls to delimit species has several practical advantages, not least the ease and economy of sound recording and analysis (Remsen 2005). Thirdly, collections in sound libraries (e.g. British Library Sound Archive, London) and on the internet (e.g. http://www.macaulaylibrary.org; http://www.xeno-canto.org) now rival museum specimen collections and online genetic databases in the diversity of taxa represented, and the depth of sampling per taxon.

Ecological divergence also provides a clue to the reproductive compatibility of taxa, particularly as hybridization between distinct ecomorphs or migratory types may lead to hybrid inviability (Price 2008). However, most ecological data may only be tangentially helpful, as many species vary geographically in their microhabitat requirements, a circumstance that low sampling effort can mask. Moreover, an ecological difference between two closely related taxa would probably covary with a morphological, acoustic or behavioural difference (Losos *et al.* 1997, Patten *et al.* 2004, McCormack & Smith 2008), so that ecology is not necessarily an independent factor. On the other hand, strict specialization in habitat preferences, along with other key ecological factors such as foraging behaviour, choice of nest-site, timing of breeding season, host-use (in brood parasites) and even migratory pattern, may have a bearing on taxonomic status (Sorenson *et al.* 2003, Friesen *et al.* 2007, Rissler & Apodaca 2007, Rolshausen *et al.* 2009).

Finally, innate behaviours unique to a taxon or group of related taxa may have taxonomic significance. For example, the split of *Hippolais opaca* from *Hippolais pallida* (with its races *elaeica*, *reiseri*

and *laeneni*), proposed in Ottosson *et al.* (2005), is reinforced by the fact that all races of *pallida* share a tail-dipping habit which *opaca* lacks. Differences in innate courtship behaviours are particularly likely to be relevant, whereas learnt or plastic behaviours, such as tameness, are less useful and should be discounted. We emphasize that this does not apply to the underlying structure of songs, reflected in basic measurements such as peak frequency and pace of notes, which have a largely genetic basis, even in oscine passerines with a high degree of song learning (Marler & Slabbekoorn 2004).

Assessing phenotypic divergence under the BSC

Museum skins cannot reveal the vocalizations, behaviours and ecological traits underlying mating preferences and zygotic compatibility in the living bird. Therefore, taxonomy traditionally rests on the evidence of phenotypic characters that play an uncertain role in reproductive isolation. The BOU guidelines responded to this issue by asserting that 'it does not matter whether or not characters used in diagnosis are relevant to the birds themselves'. This approach was theoretically sound because 'any property that provides evidence of lineage separation is relevant to inferring the boundaries and numbers of species' (de Queiroz 2007). However, we are not merely trying to identify lineages that have separated, but those that have separated in such a way that they are likely to be reproductively isolated.

From this perspective, the most appropriate taxonomic characters will be those contributing directly to reproductive isolation. A single character that makes such a contribution must inevitably have greater significance than several characters that do not. In other words, minor differences in gene sequences, or in characters like bill length and shade of plumage coloration, explicable as adaptations to local environments, are perhaps unlikely to represent barriers to interbreeding, whereas the opposite will be true of differences in characters related to courtship or species recognition (Marler 1957, Konishi 1970, Grant & Grant 1997, 2008). By the same logic, large differences in any phenotypic trait are more likely to reflect lineage separation and reproductive isolation than small differences. Previous criteria generally gloss over vocal or behavioural differences, and lump together minor and major phenotypic differences as characters with equal weighting (if they are 'diagnosable'). Our system focuses more attention on mating signals such as songs and displays, and weights all characters according to the degree of their divergence.

It is worth pointing out that even selecting target taxa for comparison raises challenges (Appendix S1). The choice is relatively straightforward when a 'species' is made up of only two disjunct subspecies. However, with an increasing number of subspecies the pattern of character distribution can become highly complex. In many polytypic taxa, subspecies vary from highly distinct to highly indistinct, so that any prospective revision of species limits requires multiple comparisons between taxa. But which taxa? – the geographically closest, the phenotypically closest, the nominate? The answer may vary on a case-by-case basis, but as a general rule we suggest that phenotypically close taxa should be compared even when geographically distant. We also suggest that – at least for conservation purposes – individual taxa be split off if they meet the criteria, rather than waiting for a full review of all subspecies.

Geographical relationships

In line with considerations outlined in the BOU guidelines, we propose that the evidence for character differentiation must increase in tandem with geographical separation. In other words, the degree of differentiation required to trigger species status must increase from parapatry through narrow hybrid zones and broad hybrid zones to allopatry. However, assigning cases to categories is not straightforward, and some practical clarifications are required.

We follow the definitions of sympatry, parapatry and allopatry provided by Futuyma and Mayer (1980):

1 'Two populations are sympatric if individuals of each are physically capable of encountering one another with moderately high frequency. Populations may be sympatric if they are ecologically segregated, as long as a fairly high proportion of each population encounters the other along ecotones'.
2 Two populations are parapatric if they occupy 'separate but adjoining regions, such that only a small fraction of individuals in each encounters the other'.

3 Two populations are allopatric if they are 'separated by uninhabited space (even if it is only a very short distance) across which migration (movement) occurs at very low frequency'.

Taxa appearing parapatric on the evidence of mapped ranges should be treated as allopatric unless their populations actually come into contact (Appendix S1). Those appearing sympatric may replace each other on elevational or ecological gradients, in which case they should be treated as parapatric if few individuals are likely to interact with heterospecifics. If contact between populations is greater, however, they may be classed as sympatric. Where habitat heterogeneity allows extensive geographical overlap without hybridization, taxa should be treated as sympatric even if their ranges appear to interdigitate in a strictly non-overlapping mosaic, so long as there is evidence that the taxa frequently come into contact. If ranges are in contact and data depth appears to be sufficient, then taxa should be treated as parapatric if records of hyridization are absent or rare.

At the junction of species ranges, stable hybrid zones are more common than true parapatry, but equally difficult to categorize. It is generally accepted that narrow hybrid zones provide better evidence of reproductive incompatibility, but how should we define a narrow zone in relation to a broad zone? Price (2008: pp. 326–328) tabulates data from 23 studies of interspecific hybrid zones varying in width from < 10 km to 1000 km, by which we can calculate a mean zone-width of 224 km. Based on these figures, we classified narrow zones as < 200 km wide, and broad zones as ≥ 200 km wide, at the maximum point.

A taxonomic scoring system

To address the inherent subjectivity of assigning taxonomic rank to allopatric, parapatric and hybridizing forms, we propose a simple point-based system whereby phenotypic differentiation between taxa is scored according to four degrees of magnitude (*minor*, *medium*, *major* and *exceptional*). These categories are defined, as far as possible, according to quantitative thresholds (see below). Overall divergence is then summed and compared with that found in irrefutable species.

We apply our criteria in two steps (Table 1). As a first step, we quantify differentiation in three main types of trait (morphology, voice and biometrics, interpretable against all four degrees of magnitude) and one subsidiary trait (ecological and behavioural characters, interpretable only against minor or medium degrees of magnitude). *Minor*, *medium*, *major* and *exceptional* differences are given scores of 1, 2, 3 and 4, respectively, and all scores are summed to generate an overall score of phenotypic divergence. As a second step, we assign taxa to one of four conditions of geographical relationship: *allopatry*, *broad hybrid zone*, *narrow hybrid zone* and *parapatry*. These are given scores of 0, 1, 2 and 3, respectively. Taxa are treated as species if their overall score reaches 7 (see Results).

Where possible, we score the strength of characters according to effect sizes computed from the means and standard deviations of sets of measurements. Effect sizes are more suitable for taxonomic appraisals than are *P*-values, which are highly correlated with sample size (Nakagawa & Cuthill 2007, Appendix S1). The fact that it is easy to achieve statistically significant differences merely by increasing sample size may lead to inappropriate taxonomic decisions (Patten & Unitt 2002). Effect sizes are most commonly presented as the Cohen's *d* statistic, which combines a measure of the magnitude of a difference with a measure of precision. On the basis of the distribution of effect sizes produced by empirical tests of divergence in undisputed species (see Results), we scored character differences with an effect size of 0.2–2 as *minor*, 2–5 as *medium*, 5–10 as *major* and > 10 as *exceptional*. This approach assumes that we can calibrate the 'significance' of effect size differences according to divergence measured across a sample of known species.

Although some plumage features, such as the width of colour patches on flight feathers, can be measured in terms of size, not all traits can be measured and converted into effect sizes using the methods outlined above. For example, it is difficult to quantify the shape of feather spots or the density of striations in a meaningful way. Similarly, colours of plumage and bare-parts can be measured using quantitative techniques, but these require sophisticated instruments, as well as complex computational processing to take into account differences in colour vision between birds and humans (e.g. Endler & Mielke 2005). Given that human vision can serve as a valid, if imperfect, proxy for avian vision (Armenta *et al.* 2008, Seddon *et al.* 2010), we have opted to rely on

Table 1. Procedure for score allocation based on five classes of taxonomic character. If the combined totals reach or exceed 7, species status is assigned. For details, see Methods.

		Magnitude (score)			
Trait type or context	Frequency of scoring	Minor (1)	Medium (2)	Major (3)	Exceptional (4)
(1) Morphology (biometrics)	Strongest increase and strongest decrease only	Effect size: 0.2–2	Effect size: 2–5	Effect size: 5–10	Effect size: >10
(2) Acoustics	Strongest temporal and spectral character only	Effect size: 0.2–2	Effect size: 2–5	Effect size: 5–10	Effect size: >10
(3) Plumage and bare parts	Three strongest characters	A slightly different wash or suffusion to all or part of any area	Distinctly different tone/shade to all or part of a significant area of feathering	Contrastingly different hue/colour to all or part of a significant area of feathering	Radically different coloration or pattern to most of plumage (striking contrast in colour, shade, shape)
(4) Ecology and behaviour	Once	Non-overlapping differences in (a) foraging/breeding habitat; (b) adaptations related to foraging/breeding; or (c) an innate habit	Non-overlapping differences in an innate courtship display	n/a	n/a
(5) Geographical relationship	n/a	Broad hybrid zone	Narrow hybrid zone	Parapatry	n/a

n/a, not applicable.

qualitative judgement for differences in colour and pattern until appropriate technology is more widely available.

We propose that visible plumage characters are categorized as follows: (i) a *minor* difference involves weak divergence in a plumage or bare part feature, i.e. a slightly different wash or suffusion to all or part of any area of feathering or bare part; (ii) a *medium* difference involves a distinctly different tone (shade: light yellow vs. dusky yellow, etc.) to all or part of a significant area of feathering (e.g. head, mantle and back; rump; wings; tail); or a strongly demarcated part of these areas (broad supercilium, breast-band, etc.) or bare part; (iii) a *major* difference involves a contrastingly different hue (colour: e.g. white/yellow; red/brown; green/blue) to all or part of a significant area of feathering (as medium above) or bare part, or the presence of an entirely different pattern (e.g. strong spotting vs. strong stripes); (iv) an *exceptional* difference involves a radically different coloration or pattern (a striking contrast of colours, shades or shapes) involving the majority of the plumage area, or any trait directly involved in courtship and mate choice.

Scoring plumage traits qualitatively in this way increases the subjectivity of our approach. However, it has the benefit of rendering the procedure both simple and inexpensive, which helps to ensure that it is open to use by anyone seeking to make species-rank assessments. Moreover, a clear scoring system minimizes subjectivity and maximizes repeatability (see below). Plumage character assessments can be made using published illustrations if necessary, although the accuracy and consistency of our method is improved with reference to museum skins.

Sampling and non-independence

Phenotypic comparisons raise many practical considerations, particularly in relation to sampling. Regardless of the target trait, or the sophistication of analysis, an inadequate or biased geographical sample will generate spurious estimates of overall divergence (Remsen 2005). For instance, the likelihood that species status is triggered will be high if characters are sampled from either end of a cline, and low if they are sampled throughout the cline. Thus, traits should be sampled from a scatter of regions, including those geographically close to the comparison taxon (Isler *et al.* 2005, Remsen 2005). For similar reasons, quantitative comparisons should be made between multiple individuals (ideally more than 10), each from the same age group, sex and race.

Another important consideration is the independence of samples. The BOU guidelines specify that morphological characters used in taxonomic assessments should be 'functionally independent'. We emphasize that functional independence not only implies that characters are used in 'separate functional contexts', but that they are not covariant, or caused by multiple phenotypic effects of a single gene (i.e. pleiotropy). For example, size-related characters, such as tarsus-length and wing-length, often covary and therefore cannot be treated independently. Similarly, covariance in colour-related traits, such as a whiter belly, broader white wingbars and a larger white rump-patch, may be driven by the same genes underlying pigmentation (e.g. reduced melanin: see Theron *et al.* 2001). In this case, related plumage features should be collapsed into a single character.

We use two methods to limit the triviality and interdependence of characters. First, we propose that taxa cannot be elevated to species rank solely on the basis of minor characters, as these can easily exceed thresholds without ever amounting to a compelling degree of differentiation; this rule applies to all four conditions of geographical relationship above. Secondly, we propose a cap on the number of scores generated within each category. Thus, total scores can only include a maximum of (i) two biometric characters (the largest increase, and the largest decrease in effect size), (ii) two vocal characters (the largest temporal, and the largest spectral effect size), (iii) three plumage characters, and (iv) one behavioural or ecological character.

Capping of biometric (i) and vocal (ii) divergence at two characters allows the maximum number of characters with the minimum degree of covariance. Note that in cases where all biometric characters increase (or decrease) in size, only one character contributes to the total score. We are able to include a third plumage character (iii) because these differences vary along multiple axes of colour and pattern and are much less prone to non-independence, particularly as non-independent plumage characters are excluded at the outset. Characters that do not vary along the same axis of size or colour are assumed to be functionally independent. Likewise, different characters thought to be used in mate choice or social signalling (e.g. song, throat-patch and tail-streamers) are treated

as independent. Characters found only in males, or only in females, are treated separately. If a trait is expressed asymmetrically in both sexes, only the largest score is calculated. We only admit one behavioural or ecological character (iv) as these are often correlated. Categories are defined in Table 1.

Setting thresholds

To produce a quantitative general estimate of the level of divergence indicative of a barrier to gene flow, we compiled a dataset of standard morphological and vocal measurements, plus scores of plumage divergence, from 58 pairs of closely related congeneric sympatric or parapatric species. These species pairs contain representatives from 29 avian families, and were selected in consultation with regional experts (see Acknowledgements). We sampled from as broad a geographical and taxonomic range as possible, while ensuring that all species pairs had similar songs and/or morphology. For a complete list of species, along with sources of specimens and sound recordings, see Appendix S2.

Biometric evidence

We collected biometric data from 53 of 58 species pairs using specimens housed at the Natural History Museum in Tring, UK, and Louisiana State University in Baton Rouge, LA, USA. Data for one further species pair were drawn from the literature (Appendix S2). Our own measurements were based on material collected from within the area of sympatry, where possible. We took standard measurements from five to 21 individuals per species (mean ± sd = 14.0 ± 2.6). We discounted body mass data as being too plastic (i.e. variable depending on time of day and season, etc.). We selected males where possible, and only used unsexed birds if they were indistinguishable from known males in size or phenotype. Dial callipers were used to measure (to the nearest 0.1 mm) (i) culmen length from tip to skull, (ii) tarsus length from where the tarsus meets the foot to the notch between the tibia and tarsus, (iii) wing length (unflattened wing chord), and (iv) tail length from the tip of the uppertail (central feathers wherever possible) to the 'point of insertion' (where the shaft disappears into the skin). We generally limited measurements to these traits, but we also used callipers to measure additional diagnostic traits (e.g. hind-claw length, bill depth or bill width) on a case-by-case basis.

Acoustic evidence

We analysed the songs of 54 of our 58 species pairs. We focused on songs rather than call notes, as songs tend to function in mate-choice and hence in reproductive isolation in birds (Collins 2004). We defined songs as territorial or advertising signals; these were generally identifiable by their complexity or stereotypy in relation to alarm or contact calls. Song recordings were compiled from the British Library Sound Archive and Macaulay Library (Cornell University), and from commercially available CDs, online sound archives and personal collections (Appendix S2). We obtained recordings from within the zone of sympatry, where possible. Final samples contained songs from two to 10 individuals per species (mean ± sd = 4.6 ± 2.1), with 1–10 songs per individual (4.6 ± 2.3). Where there was much intra- and inter-individual variation (as in many oscine species) we sampled at least six individuals per species, and at least six songs per individual. Multiple songs were often analysed from the same recording.

Using AVISOFT SASLab Pro version 4.0c (© 2002 Raimund Spect, Berlin, Germany) recordings were digitized at a sampling frequency of 44.1 kHz. We used only good quality recordings with low background noise. Amplitude was adjusted to ensure that the maximum relative amplitude of the recording was −9 dB (using Adobe AUDITION). Contrast was adjusted according to recording intensity to ensure that all elements were retained, while minimizing reverberation between elements. Song structure was then quantified using seven standard temporal (in s) and spectral (in kHz) measures:

1 total number of notes,
2 duration of song,
3 pace (number of notes divided by duration),
4 maximum frequency,
5 minimum frequency,
6 bandwidth (maximum minus minimum frequency), and
7 peak frequency (the frequency with the greatest amplitude).

These variables are easily measured from all bird songs using widely available free software (e.g. RAVEN LITE). They also show high repeatability and species specificity (J. A. Tobias *et al.* unpubl. data). Variables 1–6 were measured using on-screen cursors, and variable 7 was extracted automatically from amplitude spectra. To achieve maximum temporal resolution (1.5 ms), time features were

taken from spectrograms generated using broad-band (324 Hz) filter settings (FFT = 512). To maximize frequency resolution (43 Hz), spectral measures were taken from spectrograms produced using narrow-band (162 kHz) filter settings (FFT = 1024). As with morphology, other measurable characters can also be included on a case-by-case basis (change in pace or pitch, inter-note interval, etc.).

Several practical considerations should be borne in mind when assessing song divergence. For instance, it is essential that comparisons are made, not only with the corresponding age/sex class, but also with analogous vocalizations. For example, it is easy in poorly sampled cases to presume calls of taxon A are comparable to songs of taxon B, or primary song of taxon A to secondary song of taxon B. Doing so will result in spurious estimates of phenotypic divergence. It is also worth noting that the most diagnostic acoustic signal for analysis is not always the song. For example, reproductive isolation is sometimes maintained between sympatric or parapatric taxa with similar complex vocalizations (i.e. songs), although in these cases species often have different simple vocalizations (i.e. call notes: Martens *et al.* 2003, Tobias & Seddon 2009, Seddon & Tobias 2010). In other cases, vocal signals are sometimes near-identical between related taxa, whereas mechanically produced signals are divergent, as is the case with the calls and drums of some woodpeckers. The most diagnostic acoustic traits should be sought when selecting traits for quantification.

Testing taxonomic hypotheses with vocal data in birds is also complicated by the issue of learning, which produces complex songs, dialects and repertoires. Taxa in which learning is minimal or absent (i.e. suboscine passerines and most non-passerines) and those in which learning is widespread (i.e. oscine passerines, hummingbirds and parrots) are likely to differ in terms of individual and regional variation. Moreover, vocal divergence in learners is likely to involve many more subtle acoustic characters than vocal divergence in non-learners. For these reasons, equivalent differences in acoustic structure might reflect species status in a suboscine passerine (e.g. Isler *et al.* 1998), but nothing more than geographical dialects in an oscine passerine. Our criteria address this issue by restricting measurements to general characters, thereby limiting the influence of character triviality. Overall, relatively crude temporal or spectral characters appear to be more taxonomically informative, whereas fine-scale analyses over-emphasize divergence in oscines, particularly mimetic species (J. A. Tobias *et al.* unpubl. data; see Results).

Qualitative evidence

To produce overall divergence scores, quantitative evidence was combined with qualitative evidence in the form of plumage characters and behavioural and ecological differences. Two observers (N.J.C. and L.D.C.F.) visually scored divergence in plumage characters for all 58 species pairs, using criteria outlined above. Judgements were based on samples of museum specimens where possible (47 species), or else the colour plates from *The Handbook of the Birds of the World* (11 species). Scores were produced independently but with a high degree of congruence (Pearson's correlation: $R^2 = 0.89$). Observers then conferred and agreed on a final score (Appendix S2).

Ecological and behavioural data were compiled from relevant literature. We treated ecological characters as distinct (i.e. non-overlapping) differences in (i) foraging and/or breeding habitat preferences or (ii) adaptations related to foraging and/or breeding. We treated behavioural differences as distinctive innate habits (e.g. wing or tail movement) or displays. All ecological and behavioural distinctions were clear-cut; and because degrees of distinctiveness in these characters appear difficult to discriminate, we scored them as *minor* characters only, unless they were clearly related to courtship, in which case they were scored as *medium* characters.

Converting to effect sizes

For each species in our sample, we calculated the mean and standard deviation for the seven vocal and four biometric characters (Appendix S2). We then used an effect size calculator (widely available online) to calculate Cohen's d statistics for each species/subspecies pair and each character, as follows:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}}$$

where $\bar{x}$ = mean of species 1 and 2, s = standard deviation, and

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}$$

where n = number of individuals sampled in species 1 and 2.

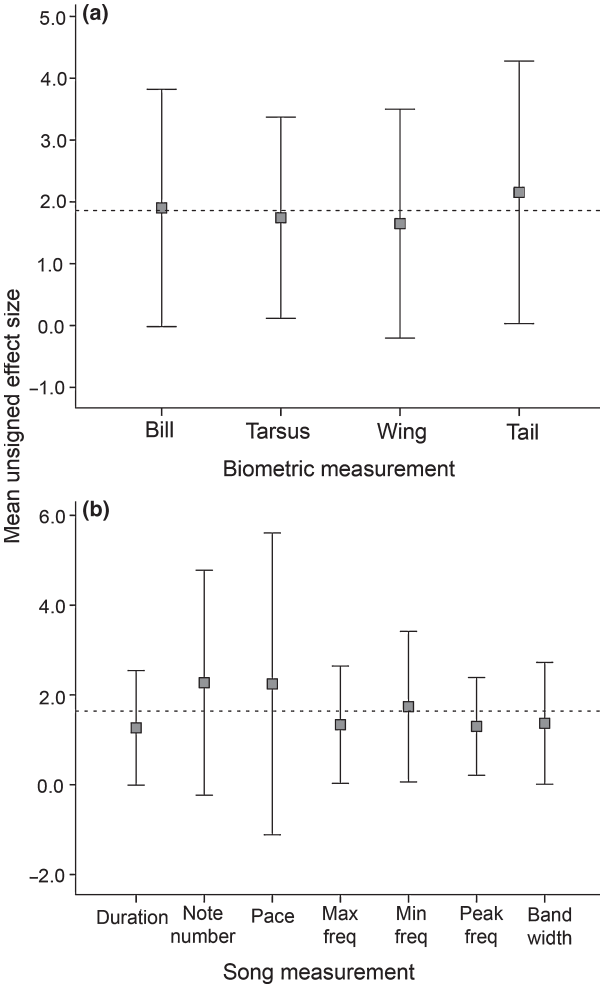

Figure 1. Mean (± sd) unsigned effect size (Cohen's *d*) across (a) four biometric and (b) seven vocal characters analysed for pairs of congeneric sympatric species (n = 53 in a; n = 54 in b; see Appendix S2). The dashed line denotes overall mean effect size across all characters within each type of trait.

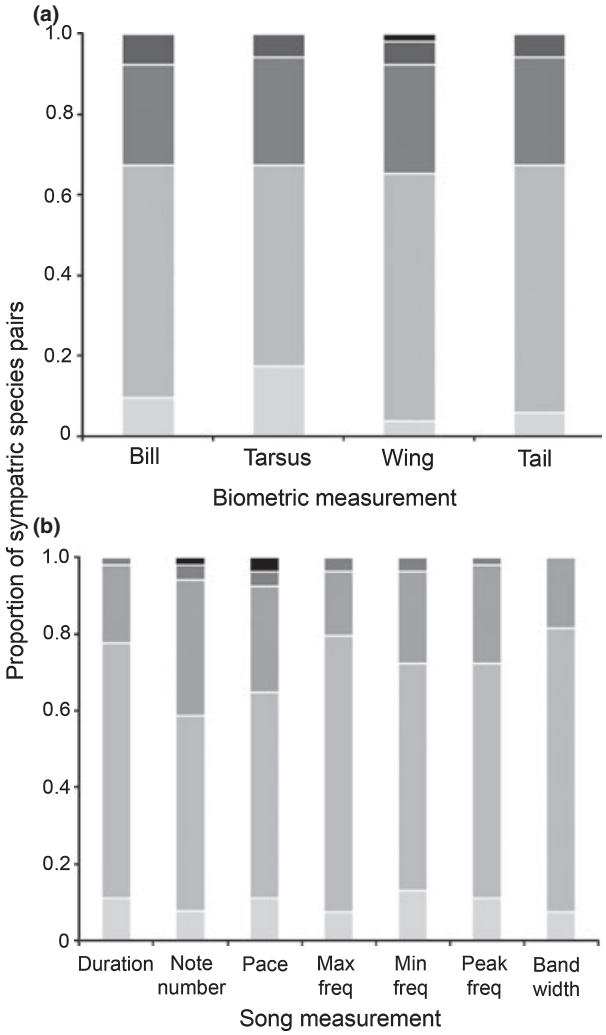

Figure 2. Proportion of sympatric species pairs with *minor*, *medium*, *major* and *exceptional* differences in (a) biometric and (b) vocal characters, with thresholds for triggering these categories set at 0.2, 2.0, 5.0 and 10.0 (Cohen's *d*), respectively. Palest bar represents traits that did not qualify (i.e. Cohen's $d < 0.2$), with four progressively darker bars indicating *minor*, *medium*, *major* and *exceptional* differences, respectively.

Comparative analyses

We used Cohen's *d* values to assess phenotypic divergence at the species level, and to set thresholds for discriminating between *minor*, *medium*, *major* and *exceptional* categories. We then assessed the effect of these thresholds on the distribution of categories across the traits and among the species pairs. Using all traits, we also explored the effect of capping the number of characters within traits on the mean and variance of total scores, and on the congruence of observer scores for plumage. We then investigated the effect of missing data by including and excluding song traits, and compared effect size distributions between oscine and suboscine species to test the extent to which our measures were influenced by song learning. Finally, we examined the relationship between divergence in morphology and song across all species pairs.

We set the threshold for triggering species status as the total capped score that resulted in 95% of sympatric taxa being assigned as species (restricting the analysis only to those species pairs with full arrays of data: biometric, vocal, visual and ecological). We then used data from a sample of 23 pairs of subspecies drawn from the Western Palaearctic avifauna (Appendix S2) to compare taxonomic

recommendations based on our approach with those of the BOU guidelines. Data were collected from 'nearest neighbour' taxa, where possible including the nominate subspecies. Using specimens housed in the Natural History Museum, UK, we compiled data on biometrics from four to 15 individuals per subspecies (mean = 14.3 ± 1.9). These data were used to generate Cohen's *d* values, which in turn were converted to character magnitudes. For each pair of taxa, N.J.C. and L.D.C.F. generated a list of diagnostic plumage and bare-part characters from Cramp (1977–1994), and then used museum specimens to score these differences as *minor*, *medium*, *major* and *exceptional*, as above. Biometric and plumage scores were converted into total capped scores, and taxonomic status was assigned according to our criteria. Using the same literature and specimen data, we then assigned species limits according to the BOU guidelines (Table S1). We discounted acoustic data from both analyses because (i) they were only available for a subset of taxa, and (ii) the BOU guidelines do not include information about dealing with acoustic data, making it impossible to standardize acoustic comparisons. Comparisons on the basis of incomplete datasets are not full tests of our criteria. However, they remain valid and highly informative about the differences between the proposed systems (see Discussion). We also note that it was difficult to invoke the 'divergence evaluation test' under the BOU guidelines, leading to uncertainty regarding the taxonomic rank supported.

RESULTS

Comparing across all species pairs, our data revealed that vocal characters had greater variability but slightly lower mean levels of divergence than biometric data. Specifically, the mean ± se unsigned effect size across all four biometric characters was 1.86 ± 1.88 (range = 0–10.73; $n = 53$; Fig. 1a), whereas the equivalent mean unsigned effect size across all vocal characters was 1.64 ± 0.92 (0–18.0; $n = 54$; Fig. 1b). However, mean effect sizes were fairly consistent across all biometric and vocal characters, being close to 2.0 in all cases (Fig. 1). On the basis of this result we set the threshold for triggering a *medium* difference as Cohen's $d = 2.0$. In addition, the thresholds for *major* and *exceptional* differences were arbitrarily set at $d = 5.0$ and $d = 10.0$, respectively, within the tail of the effect size distribution generated by our sample (Fig. 2). Finally, we treated any effect size larger than 0.2 as biologically relevant, following Cohen (1988), and hence the threshold for a minor difference was set at $d = 0.2$.

When these four thresholds were applied to biometric characters we found that 58.0 ± 5.9% of pairs had a *minor* difference, 26.0 ± 1.1% had a *medium* difference, 6.0 ± 1.1% had a *major* difference and 0.5 ± 1.1% had an *exceptional* difference ($n = 53$; Fig. 2a). We also found that a similar proportion of pairs had vocal differences in each of these categories: 63.0 ± 9.0% of pairs had a *minor* difference, 24.0 ± 6.0% had a *medium* difference, 3.0 ± 1.0% had a *major* difference and 1.0 ± 1.0% had an *exceptional* difference ($n = 54$; Fig. 2b). Visual assessment of plumage characters generated a slightly different category ratio, with more characters detected overall, higher numbers of medium and major characters, but very few exceptional characters. All but one species pair (98.3%) had at least one *minor* difference in plumage, 47 pairs (81.0%) had at least one *medium* difference, 18 pairs (31%) had at least one *major* difference and only one pair had an *exceptional* difference.

By applying our thresholds and summing scores across all traits (i.e. no capping), we found that species pairs had high and variable total scores of phenotypic difference: 18.7 ± 3.9 (range = 10–28; $n = 49$). In contrast, capping the number of characters within each trait produced a lower, less variable total score of 10.4 ± 2.0 (5–14). Focusing exclusively on qualitative plumage assessments showed that species pairs differed substantially by plumage characters, with a mean uncapped score of 5.28 ± 2.57 (0–13), and a mean capped score of 4.66 ± 1.73 (0–8; Appendix S2). Note that the application of capping to plumage characters again resulted in reduced variance. Moreover, it also resulted in greater agreement between observers during plumage divergence assessments (without capping: $R^2 = 0.78$; with capping: $R^2 = 0.89$).

Using capped data, and restricting the analysis to biometric, visual and ecological data, the mean total divergence score was 6.87 ± 1.71 ($n = 53$ pairs; Fig. 3a); when song was added, this increased to 10.4 ± 2.0 ($n = 49$; Fig. 3b). When we pooled capped data in this way, we found that 95% of all sympatric taxa were correctly classified as species by a cut-off at 7.0 (Fig. 3b). We conclude that a total capped score of 7.0 is an appropriate threshold for triggering species status according to our criteria.

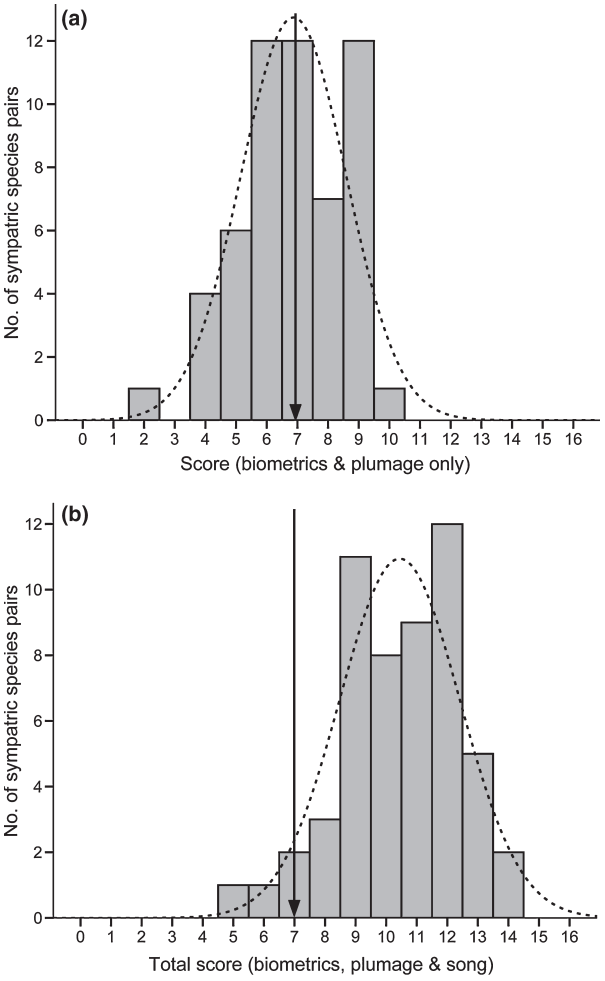

Figure 3. Overall levels of phenotypic divergence across sympatric species pairs. Shown are the distributions of scores summed across (a) biometrics and plumage only (n = 53), and (b) biometrics, plumage and song (n = 49). Depicted are the normal distribution curve (dashed black line) and the threshold (7.0) for triggering species status (vertical arrow). Note that when using this threshold 95% of pairs of taxa are correctly classified as species.

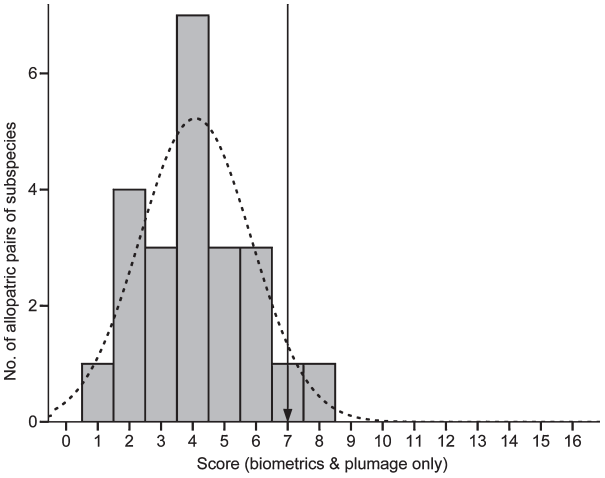

Figure 4. Distribution of total scores for allopatric pairs of subspecies in European birds. Note that 21 of 23 pairs of taxa retain subspecies status when the threshold for triggering species status is set at 7.0, although the addition of vocal data may increase this proportion.

When we tested this system, along with the BOU guidelines for comparison, against 23 pairs of European subspecies, we found that the two approaches produced different taxonomic recommendations (Appendix S2). Under our system, phenotypic divergence between subspecies was lower than that found between sympatric species: Cohen's d values ranged between 0.01 and 3.69 (mean = 1.16 ± 0.91) for biometric characters, and total uncapped plumage scores between 0 and 6.0 (2.0 ± 1.5). The mean total score summed across biometric and plumage characters (with capping) for allopatric subspecies was 4.1 ± 1.8 (Fig. 4), rather than 7.0. Only two pairs of taxa (9%) qualify for species status under our criteria: *Oenanthe oenanthe oenanthe* and *Oenanthe oenanthe seebohmi*; and *Rhodopechys sanguinea sanguinea* and *Rhodopechys sanguinea aliena* (Appendix S2). Meanwhile, 17 pairs of taxa (74%) convert to species (or probable species) under the BOU guidelines as we interpret them (Appendix S2). Under both systems, the proportions of taxa considered species would probably increase with the addition of vocal data. However, as vocal differences between most subspecies are likely to be far smaller than those between sympatric species, we believe that the addition of vocal characters would still produce fewer taxonomic changes under our system than the BOU guidelines.

Further analysis of phenotypic variation in our species-level dataset revealed no significant difference between oscine and suboscine passerines in the mean unsigned effect size for any vocal character (unpaired t-test; $0.22 < t < 1.75$, $0.087 < P < 0.827$, n = 33 pairs of oscines, 17 pairs of suboscines). This suggests that our method is robust to variation in song learning and song complexity. We also found a negative correlation between total uncapped scores when comparing plumage with song scores (Spearman's rho = –0.33, P = 0.015, n = 54 species pairs). This reflected the fact that some species with divergent morphology had

similar songs (e.g. *Neocossyphus poensis* and *Neocossyphus rufus*), whereas other species with divergent songs had very similar morphology (e.g. *Empidonax alnorum* and *Empidonax traillii*). This result does not correct for phylogenetic effects, and so has to be interpreted with caution. Nonetheless, it suggests that increased divergence in acoustic characters was associated with reduced divergence in visual characters.

DISCUSSION

The taxonomic methods outlined here are similar to previous approaches in that decisions are guided by direct reference to known sympatric species. However, the novel addition of a quantitative scoring system allows us to make this calibration more accurate and explicit. Allopatric and parapatric species are treated as hypotheses which can be tested. Our system is therefore analogous to a hypothetico-deductive framework. In this case, the null hypothesis is that different populations are representatives of a single species, and the significance level of empirical tests is set by the degree of divergence in undisputed species. Unless an allopatric form passes these quantitative thresholds, we conclude that it should be left as a subspecies of a wider ranging taxon until better evidence – genetic, ecological or behavioural – permits rejection of the null hypothesis.

Our results suggest that this approach is workable and relatively consistent. In particular, they confirm that the exclusion of trivial characters, and the capping of character number, limits the accumulation of scores based on minor characters. It also shifts focus to the largest differences between taxa, which are more likely to be associated with reproductive isolation. Testing the criteria against European subspecies limits suggests that our system produces results that align with general standards in Europe concerning biological species. This contrasts with the application of the BOU guidelines, which prove to be more ambiguous and lead to many more taxonomic changes. Although this finding may suggest that our thresholds are designed to maintain the taxonomic status quo, it should be borne in mind that vocal data were not included in the analyses, and therefore the addition of these data may trigger species status for some vocally divergent subspecies (e.g. *Sylvia sarda sarda* and *Sylvia sarda balearica*) under our criteria (Fig. 3). More importantly, Collar (2006a,b, 2007) tested an earlier version of the same criteria against a larger suite of Asian taxa, and found that it allowed the separation of 48 'new' species, of which three have already been supported by independent molecular work (Li Shou-Hsien *et al.* 2006, Feinstein *et al.* 2008). Overall, we conclude that our scoring system can deliver robust and consistent taxonomic change within the conceptual framework of the BSC.

Two other inferences can be made from our results. First, they suggest that the methods for capping character number – which were introduced to reduce the difficulty posed by non-independent characters – have the added benefit of reducing uncertainty and maximizing repeatability. Secondly, they reveal an inverse correlation between divergence in plumage and morphology on the one hand, and song on the other, supporting the idea that there is an evolutionary trade-off between different forms of costly traits (Badyaev *et al.* 2002). In other words, a taxonomy based on morphological and plumage characters alone will tend to overlook cryptic species whose divergence is biased towards acoustic characters. This highlights the importance of including both morphology and song when assessing species limits in birds.

Strengths and limitations of the system proposed

Species boundaries are inherently 'fuzzy', and it is therefore easy to find fault with any species concept and any system of rules for delimiting species (Hey 2001, Mallet 2008, Joseph & Omland 2009). In effect, an operational approach will always require researchers to make qualitative judgements, and no system involving the subdivision of a continuous process can hope to be fully objective (Hey 2001, 2006, Sites & Marshall 2004). However, this becomes less of an obstacle once it is recognized that the priority is not perfect objectivity but effective compromise (Hey *et al.* 2003, Winker *et al.* 2007). We accept that our system can neither cover every context nor cater for all tastes, but we hope that it may help promote accuracy, stability and consistency in avian taxonomy.

As with any system partially based on qualitative judgement, one of the main difficulties faced by our method is subjectivity (Peterson & Moyle 2008). It could be argued, for example, that much depends on the distinction between minor and medium characters, opening the way for slight

ambiguities to blur the boundary between species and subspecies. Nonetheless, difficulties of this type are unavoidable in any taxonomic appraisal (Sites & Marshall 2004), and we minimize their impact by introducing explicit quantitative or verbal thresholds. We argue that these thresholds lend our approach a higher degree of objectivity than achieved by the PSC, which – because of the lack of constraints to character triviality – merely shifts the problem of arbitrariness to a finer scale (Collar 1997, Johnson *et al.* 1999, Coyne & Orr 2004, Winker *et al.* 2007, Price 2008).

The dependence of our criteria on a fixed overall threshold (seven character points) may be interpreted as a procedural weakness. Can the system work if *Acrocephalus* warblers and *Empidonax* flycatchers choose their mates differently from ducks or parrots? Will it be effective if a character of given magnitude is relevant to one group but not to another? Our analysis suggests that the criteria largely overcome this problem by combining datasets, because divergence at the species level is fairly consistent when the full array of traits is taken into account (Fig. 3). In other words, divergence in ducks and parrots may occur in visual traits, while that in warblers and flycatchers may occur in vocal and behavioural traits, but species status is reflected in a relatively uniform shift in overall phenotype. Thus, on the one hand, our data support the observation of Mayr and Gilliard (1952, p. 334) that 'reproductive isolation and morphological divergence are not closely correlated'. On the other hand, they indicate that divergence summed across morphology, plumage, behaviour, ecology and acoustic signals is much more likely to be correlated with reproductive isolation.

This finding validates the use of a fixed threshold, and also argues against the use of a two-tier system, i.e. one threshold for phenotypic data and another (raised) threshold for phenotypic plus vocal data. Although fewer cases are correctly classified as species if our single-threshold analysis contains only non-vocal traits (Fig. 3), we do not see this as a shortcoming. Under a two-tier system there would be a temptation to opt for the lower threshold by eliminating vocal data, which may lead to cryptic species being overlooked. A one-tier system has the benefit of encouraging consideration of multiple characters, an approach that has been widely recommended (e.g. Edwards *et al.* 2005, Yoder *et al.* 2005, de Queiroz 2007, Alström *et al.* 2008b, Leaché *et al.* 2009). If a one-tier system classifies forms as species on the basis of morphological and ecological characters alone, then vocal analyses are not required. However, if it classifies them as subspecies, the result can be further tested by adding vocal data.

Although any fixed threshold will erroneously classify some taxa as species, and others as subspecies, a key strength of our criteria is that they correctly classify the vast majority of taxa if datasets are combined. We also argue that the use of a fixed threshold brings a number of advantages, not least of which are stability and clarity. It renders our system easily understood and communicated, and simple enough to be used as a taxonomic rule-of-thumb. It also means that all steps are easily tracked and reported, improving the transparency of taxonomic decisions. Stability, clarity and transparency may not be widely considered core scientific goals, but this view overlooks the great importance of standardized taxonomy as the bedrock of conservation and policy (Ryder 1986, Collar 1997, Mace 2004, Garnett & Christidis 2007), as well as for disciplines such as macroecology that use species as units of comparison (Isaac *et al.* 2004).

To bring rigour to taxonomic assessments, we use as our benchmark for species status the degree of divergence between sympatric or parapatric congeners. Our method will therefore tend to generate phenotypically distinctive species. Indeed, the development of these criteria was motivated by widespread misgivings about the subjective determination and proliferation of undistinctive PSC or molecular-based species (see Collar 1997, Johnson *et al.* 1999). Our system counters this trend, and goes some way to answering the call for taxonomic splits to be accompanied by 'sufficient evidence that morphological, ecological, behavioural, and genetic differences between the two forms are of a magnitude that would merit specific rank in closely related sympatric forms' (Meiri & Mace 2007). In adopting these standards, our criteria will classify a proportion of PSC species (Cracraft 1983), 'evolutionarily significant units' (Moritz 1994), and 'independent evolutionary trajectories' (Peterson & Moyle 2008), as intraspecific variation. However, we see this as a strength, because it ensures that 'species' meet a consistent and significant level of phenotypic divergence.

The above points deal mainly with strategies to forestall the over-splitting of biodiversity into phylospecies. However, our system is also designed

to reduce the over-lumping of polytypic species in poorly known regions. Like Helbig *et al.* (2002), we accept taxa connected by stable hybrid zones as species, as long as phenotypic divergence outside the hybrid zone meets stipulated levels. Note that we have established these levels by comparison with divergence in species-pairs occurring not only sympatrically, but also across hybrid zones. As species linked by hybrid zones tend to be less divergent than sympatric species (Price 2008), the threshold for species status is reduced, making our system sensitive enough to detect cryptic diversity.

Future directions

The criteria presented here require further testing against widely accepted species and subspecies limits. They should be viewed as a baseline which can be modified according to new information, including their performance in specific case studies. To point the way forward, we draw a parallel with the IUCN Red List and its categories of conservation status (IUCN Species Survival Commission 2001). The Red List uses arbitrary quantitative thresholds to assign species to categories, and it therefore faces many of the challenges described above. Nonetheless, it has proved to be extremely useful, and is growing more robust over time with the refinement of criteria (Rodrigues *et al.* 2006). Its success is founded on two decades of vigorous and constructive criticism, along with a pragmatic acceptance that its advantages outweigh its drawbacks.

Like the Red List process, our method involves assigning populations to categories partially on the basis of quantitative data. For some characters, this means generating effect sizes, converting effect sizes to scores and then applying scores to a threshold. This three-step process is necessary because many characters (plumage, behaviour, ecology) cannot yet be converted into effect sizes, as they are scored visually or from the literature. One possible improvement to the procedure involves adopting a quantitative approach to the measurement of colour, hue and pattern differences, perhaps using spectrophotometry. By this route, all vocal and visual characters would generate effect sizes, which could then be added together to produce a single 'test statistic' and applied directly to an effect-size threshold. This would eliminate the need for converting effect sizes to scores, reducing the process to two steps. Refinements of this kind should be explored, but we caution against making the system so technical and costly that it is off-putting to those who are likely to find it the most useful, i.e. taxonomists and conservationists. We also feel that the score-based system will need to be retained for use in data-poor scenarios.

Another possible improvement is the use of genetic evidence (Winker 2009). Thresholds of genetic divergence could easily be converted to scores and incorporated into our system along with minor adjustments to the seven-point system. A simple threshold such as 4% divergence in mtDNA coding regions may capture the majority of allopatric species as there is evidence that genomic incompatibilities tend to build up after 2 million years of isolation (Price & Bouvier 2002). However, most evidence suggests that the link between genetic divergence and reproductive isolation is extremely complex, and that 4% mtDNA divergence is an unreliable threshold (see Payne & Sorenson 2007). We therefore hesitate to propose any cut-offs at this stage, and suggest that a thorough survey of molecular divergence estimates in accepted species (either sympatric or parapatric) is required to assess the feasibility of applying genetic criteria to allopatric forms.

Any such review would need to consider a range of issues. At a practical level, incomplete sampling of localities may lead to errors and exaggerations in divergence estimates, whereas extensive sampling may uncover phylogeographical structure so complex that it is difficult to interpret in terms of species limits (e.g. Goldstein *et al.* 2000, Marks *et al.* 2002, Cadena *et al.* 2007, Nyári 2007, Smith *et al.* 2007, Miller *et al.* 2008). Measurement and comparison of genetic divergence across taxa and studies must also account for variation in evolutionary rate within and between genes. Ideally, multiple loci and individuals should be sampled across each taxon's range, including contact zones, with sampling depth standardized across species. When these issues are surmounted, molecular analyses will doubtless play an increasingly central role in global taxonomic revisions.

CONCLUSION

Disagreement about what constitutes a species is so pervasive that some authors view current taxonomic assessments of species totals as gross underestimates (e.g. Peterson 1998, Sangster 2000b) whereas others fear that they are overestimates (e.g. Chaitra *et al.* 2004, Isaac *et al.* 2004, Meiri &

Mace 2007). According to Brookfield (2002), this aspect of the species problem 'is not a scientific problem at all, merely one about choosing and consistently applying a convention about how we use a word'. We have addressed this issue at an operational level by developing a system of standardized criteria for species delimitation. It has been developed for use in birds, but can be easily adapted for use in many other taxa. Moreover, it can be applied more broadly to the study of evolutionary questions, including spatial variation, phenotypic plasticity and sexual dichromatism.

To some, the system we have proposed will seem unacceptably simplistic and subjective. Others will think it excessively complicated. However, we believe it maximizes objectivity and procedural consistency. It 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 evaluated 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 (e.g. Gamauf *et al.* 2005, Phillimore *et al.* 2008), and to produce a global taxonomy of comparable species units.

This paper grew from the work of BirdLife International's Taxonomic Working Group, which over the last decade has implemented taxonomic decisions with the aim of establishing a standardized global list of avian species. Title, content and authorship have evolved as with any work in progress, and we hope to be forgiven for advance publishing of taxonomic matters relating to highly threatened regions based on earlier versions of the criteria. We are especially indebted to Per Alström, Jim Mallet, Trevor Price, Michael Sorenson and Kevin Winker for valuable comments on earlier versions of the manuscript, and to many colleagues who offered assistance or shared their views on species concepts and species delimitation, including Paul Andrew, Axel Bräunlich, Leon Bennun, Stuart Butchart, John Croxall, Robert Dowsett, Françoise Dowsett-Lemaire, Stephen Garnett, Jamie Gilardi, Jürgen Haffer, Leo Joseph, Ian Newton, Stephen Parry, Ben Phalan, Robert Prŷs-Jones, Pamela Rasmussen and Alison Stattersfield. For access to study skins, we are grateful to Robert Prŷs-Jones and colleagues of the Natural History Museum, Tring, and James Van Remsen, Jr, of the Museum of Natural Science, Lousiana State University, USA. We thank Marina Amaral for assistance with acoustic analyses, and the following regional experts for providing recordings or suggestions for suitable congeneric sympatric/allopatric species-pairs: Callan Cohen (South Africa), Mario Cohn-Haft (South America), Françoise Dowsett-Lemaire (West and Central Africa), Guy Dutson (New Guinea), John Fitzpatrick (North America), Pete Leonard (East Africa), Pamela Rasmussen (Indian subcontinent), Glenn-Peter Sætre (Europe), and David Stewart and Richard Thomas (Australia). Thanks also to the Macaulay Library (Mike Andersen) and National Sound Archive (Cheryl Tipp) for providing many of the recordings.

REFERENCES

Abbott, C.L. & Double, M.C. 2003. Phylogeography of Shy and White-capped Albatrosses inferred from mitochondrial DNA sequences: implications for population history and taxonomy. *Mol. Ecol.* **12**: 2747–2758.

Agapow, P.M. 2005. Species: demarcation and diversity. In Purvis, A., Gittleman, J.L. & Brooks, T. (eds) *Phylogeny and Conservation*: 57–75. Cambridge: Cambridge University Press.

Agapow, P.M., Bininda-Emonds, O.R.P., Crandall, K.A., Gittleman, J.L., Mace, G.M., Marshall, J.C. & Purvis, A. 2004. The impact of species concept on biodiversity studies. *Q. Rev. Biol.* **79**: 161–179.

Alström, P. & Ranft, R. 2003. The use of sounds in avian systematics, and the importance of bird sound archives. *Bull. Br. Orn. Club* **123A**(Suppl.): 114–135.

Alström, P., Olsson, U., Lei, F., Wang, H.-T., Gao, W. & Sundberg, P. 2008a. Phylogeny and classification of the Old World Emberizini (Aves, Passeriformes). *Mol. Phylogenet. Evol.* **47**: 960–973.

Alström, P., Rasmussen, P.C., Olsson, U. & Sundberg, P. 2008b. Species delimitation based on multiple criteria: the Spotted Bush Warbler *Bradypterus thoracicus* complex (Aves: Megaluridae). *Biol. J. Linn. Soc.* **154**: 291–307.

Armenta, J.K., Dunn, P.O. & Whittingham, L.A. 2008. Quantifying avian sexual dichromatism: a comparison of methods. *J. Exp. Biol.* **211**: 2423–2430.

Badyaev, A.V., Hill, G.E. & Weckworth, B.V. 2002. Species divergence in sexually selected traits: increase in song elaboration is related to decrease in plumage ornamentation in finches. *Evolution* **56**: 412–419.

Baker, A.J. & Boylan, J.T. 1999. Singing behaviour, mating associations and reproductive success in a population of hybridizing lazuli and indigo buntings. *Condor* **101**: 493–504.

Ballard, J.W.O. & Rand, D.M. 2005. The population biology of mitochondrial DNA and its phylogenetic implications. *Annu. Rev. Ecol. Syst.* **36**: 621–642.

Benkman, C.W., Smith, J.W., Keenan, P.C., Parchman, T.L. & Santisteban, L. 2009. A new species of the Red Crossbill (Fringillidae: *Loxia*) from Idaho. *Condor* **111**: 169–176.

Brambilla, M., Vitulano, S., Spina, F., Baccetti, N., Gargallo, G., Fabbri, E., Guidali, F. & Randi, E. 2009. A molecular phylogeny of the *Sylvia cantillans* complex: cryptic species within the Mediterranean basin. *Mol. Phylogenet. Evol.* **48**: 461–472.

Braun, M.J., Isler, M.L., Isler, P.R., Bates, J.M. & Robbins, M.B. 2005. Avian speciation in the Pantepui: the case of the Roraiman Antbird (*Percnostola* [*Schistocichla*] '*leucostigma*' *saturata*). *Condor* **107**: 327–341.

Brelsford, A. & Irwin, D.E. 2009. Incipient speciation despite little assortative mating: the Yellow-rumped Warbler hybrid zone. *Evolution* **63**: 3050–3060.

Brookfield, J. 2002. [Review of] 'Genes, categories and species – the evolutionary and cognitive causes of the species problem'. *Genet. Res.* **79**: 107–108.

Brown, W.L. & Wilson, E.O. 1956. Character displacement. *Syst. Zool.* **5**: 49–64.

Cadena, C.D. & Cuervo, A.M. 2010. Molecules, ecology, morphology and songs in concert: how many species is *Arremon torquatus* (Aves: Emberizidae)? *Biol. J. Linn. Soc.* **99**: 152–176.

Cadena, C.D., Klicka, J. & Ricklefs, R.E. 2007. Evolutionary differentiation in the Neotropical mountains: molecular phylogenetics and phylogeography of *Buarremon* brush-finches (Aves, Emberizidae). *Mol. Phylogenet. Evol.* **44**: 993–1016.

Carling, M.D. & Brumfield, R.T. 2008. Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in *Passerina* buntings. *Genetics* **178**: 363–377.

Catchpole, C.K. & Slater, P.J.B. 1995. *Bird Song: Biological Themes and Variations*. Cambridge: Cambridge University Press.

Chaitra, M.S., Vasudevan, K. & Shanker, K. 2004. The biodiversity bandwagon: the splitters have it. *Curr. Sci.* **86**: 897–899.

Chesser, R.T., Barker, F.K. & Brumfield, R.T. 2007. Fourfold polyphyly of the genus formerly known as *Upucerthia*, with notes on the systematics and evolution of the avian subfamily Furnariinae. *Mol. Phylogenet. Evol.* **44**: 1320–1332.

Cibois, A., Thibault, J.-C. & Pasquet, E. 2007. Uniform phenotype conceals double colonization by reed-warblers of a remote Pacific archipelago. *J. Biogeogr.* **34**: 1150–1166.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Collar, N.J. 1997. Taxonomy and conservation: chicken and egg. *Bull. Br. Orn. Club* **117**: 122–136.

Collar, N.J. 2003. How many bird species are there in Asia? *Oriental Bird Club Bull.* **38**: 20–30.

Collar, N.J. 2006a. A partial revision of the Asian babblers (Timaliidae). *Forktail* **22**: 85–112.

Collar, N.J. 2006b. A taxonomic reappraisal of the Black-browed Barbet *Megalaima oorti*. *Forktail* **22**: 170–173.

Collar, N.J. 2007. Taxonomic notes on some insular *Loriculus* hanging-parrots. *Bull. Br. Orn. Club* **127**: 97–107.

Collins, S. 2004. Vocal fighting and flirting: the functions of birdsong. In Marler, P. & Slabbekoorn, H. (eds) *Nature's Music: The Science of Birdsong*: 39–79. London: Elsevier.

Coyne, J.A. & Orr, H.A. 2004. *Speciation*. Sunderland, MA: Sinauer Associates.

Cracraft, J. 1983. Species concepts and speciation analysis. *Curr. Ornithol.* **1**: 159–187.

Cracraft, J. 1989. Speciation and its ontology: the empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. In Otte, D. & Endler, J.A. (eds) *Speciation and its Consequences*: 28–59. Sunderland, MA: Sinauer Associates.

Cramp, S.E. 1977–1994. *Handbook of the Birds of the Western Palearctic*, Vols 1–9. Oxford: Oxford University Press.

Crandall, K., Bininda-Emonds, O., Mace, G. & Wayne, R. 2000. Considering evolutionary processes in conservation biology. *Trends Ecol. Evol.* **15**: 290–295.

Dávalos, L.M. & Porzecanski, A.L. 2009. Accounting for molecular stochasticity in systematic revisions: species limits and phylogeny of *Paroaria*. *Mol. Phylogenet. Evol.* **53**: 234–248.

DeSalle, R., Egan, M.G. & Siddall, M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Phil. Trans. R. Soc. Lond. B* **360**: 1905–1916.

Dietzen, C., Garcia-del-Rey, E., Delgado Castro, G. & Wink, M. 2008. Phylogeography of the Blue Tit (*Parus teneriffae*-group) on the Canary Islands based on mitochondrial DNA sequence data and morphometrics. *J. Ornithol.* **149**: 1–12.

Dobzhansky, T. 1937. *Genetics and the Origin of Species*. New York: Columbia University Press.

Donoghue, M.J. 1985. A critique of the biological species concept and recommendations for a phylogenetic alternative. *Bryologist* **88**: 172–181.

Edwards, S.V., Kingan, S.B., Calkins, J.D., Balakrishnan, C.N., Jennings, W.B., Swanson, W.J. & Sorenson, M.D. 2005. Speciation in birds: genes, geography, and sexual selection. *Proc. Natl Acad. Sci. USA* **102**: 6550–6557.

Elias, M., Hill, R.I., Willmott, K., Dasmahapatra, K., Brower, A.V.Z., Mallet, J. & Jiggins, C.D. 2008. Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proc. R. Soc. Lond. B* **274**: 2881–2889.

Endler, J.A. & Mielke, P.W. 2005. Comparing entire colour patterns as birds see them. *Biol. J. Linn. Soc.* **864**: 405–431.

Feinstein, J., Xiaojun, Y. & Shou-Hsien, L. 2008. Molecular systematics and historical biogeography of the Black-browed Barbet species complex (*Megalaima oorti*). *Ibis* **150**: 40–49.

Frey, J.K. 1993. Modes of peripheral isolate formation and speciation. *Syst. Biol.* **42**: 373–381.

Friesen, V.L., Smith, A.L., Gomez-Diaz, E., Bolton, M., Furness, R.W., Gonzalez-Solis, J. & Monteiro, L.R. 2007. Sympatric speciation by allochrony in a seabird. *Proc. Natl Acad. Sci. USA* **104**: 18589–18594.

Funk, D.J. & Omland, K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Syst.* **34**: 397–423.

Futuyma, D.J. & Mayer, G.C. 1980. Non-allopatric speciation in animals. *Syst. Zool.* **29**: 254–271.

Gamauf, A., Gjershaug, J.-O., Røv, N., Kvaløy, K. & Haring, E. 2005. Species or subspecies? The dilemma of taxonomic ranking of some South-East Asian hawk-eagles (genus *Spizaetus*). *Bird Conserv. Int.* **15**: 99–117.

García-Moreno, J. & Fjeldså, J. 1999. Re-evaluation of species limits in the genus *Atlapetes* based on mtDNA sequence data. *Ibis* **141**: 199–207.

Garnett, S.T. & Christidis, L. 2007. Implications of changing species definitions for conservation purposes. *Bird Conserv. Int.* **17**: 187–195.

Goldstein, P.Z., DeSalle, R., Amato, G. & Vogler, A.P. 2000. Conservation genetics at the species boundary. *Conserv. Biol.* **14**: 120–131.

Grant, P.R. & Grant, B.R. 1992. Hybridization in bird species. *Science* **256**: 193–197.

Grant, P.R. & Grant, B.R. 1997. Hybridization, sexual imprinting and mate choice. *Am. Nat.* **149**: 1–28.

Grant, P.R. & Grant, B.R. 2008. *How and Why Species Multiply: The Radiation of Darwin's Finches*. Princeton: Princeton University Press.

Hackett, S.J., Kimball, R.T., Reddy, S., Bowie, R.C.K., Braun, E.L., Braun, M.J., Chojnowski, J.L., Cox, W.A., Han, K.-L., Harshman, J., Huddleston, C.J., Marks, B.D., Miglia, K.J., Moore, W.S., Sheldon, F.H., Steadman, D.W., Witt, C.C. & Yuri, T. 2008. A phylogenomic study of birds reveals their evolutionary history. *Science* **320**: 1763–1768.

Harrison, R.G. 1998. Linking evolutionary patterns and processes: the relevance of species concepts for the study of speciation. In Howard, D.J. & Berlocher, S.H. (eds) *Endless Forms: Species and Speciation*: 19–31. Oxford: Oxford University Press.

Hebert, P.D.N., Stoekle, M.Y., Zemlak, T.S. & Francis, C.M. 2004. Identification of birds through DNA barcodes. *PLoS Biol.* **2**: 1657–1663.

Hedin, M.C. 1997. Speciational history in a diverse clade of habitat-specialized spiders (Araneae: Nesticidae: *Nesticus*): inferences from geographic-based sampling. *Evolution* **51**: 1927–1943.

Helbig, A.J., Seibold, I., Martens, J. & Wink, M. 1995. Genetic differentiation and phylogenetic relationships of Bonelli's Warbler *Phylloscopus bonelli* and Green Warbler *P. nitidus*. *J. Avian Biol.* **26**: 139–153.

Helbig, A.J., Knox, A.G., Parkin, D.T., Sangster, G. & Collinson, M. 2002. Guidelines for assigning species rank. *Ibis* **144**: 518–525.

Hendry, A.P., Vamosi, S.M., Latham, S.J., Heilbuth, J.C. & Day, T. 2000. Questioning species realities. *Conserv. Genet.* **1**: 67–76.

Hey, J. 2001. The mind of the species problem. *Trends Ecol. Evol.* **16**: 326–329.

Hey, J. 2006. On the failure of modern species concepts. *Trends Ecol. Evol.* **32**: 447–450.

Hey, J., Waples, R.S., Arnold, M.L., Butlin, R.K. & Harrison, R.G. 2003. Understanding and confronting species uncertainty in biology and conservation. *Trends Ecol. Evol.* **18**: 597–603.

Irwin, D.E., Rubtsov, A.S. & Panov, E.N. 2009. Mitochondrial introgression and replacement between Yellowhammers (*Emberiza citrinella*) and Pine Buntings (*E. leucocephalos*; Aves, Passeriformes). *Biol. J. Linn. Soc.* **98**: 422–438.

Isaac, N.J.B., Mallet, J. & Mace, G.M. 2004. Taxonomic inflation: its effect on macroecology and conservation. *Trends Ecol. Evol.* **19**: 464–469.

Isler, M.L., Isler, P.R. & Whitney, B.M. 1998. Use of vocalizations to establish species limits in antbirds (Passeriformes; Thamnophilidae). *Auk* **115**: 577–590.

Isler, M.L., Isler, P.R. & Whitney, B.M. 1999. Species limits in antbirds (Passeriformes; Thamnophilidae): the *Myrmotherula surinamensis* complex. *Auk* **116**: 83–96.

Isler, M.L., Isler, P.R. & Brumfield, R.T. 2005. Clinal variation in vocalizations of an antbird (Thamnophilidae) and implications for defining species limits. *Auk* **122**: 433–444.

Isler, M.L., Isler, P.R. & Whitney, B.M. 2007. Species limits in antbirds (Thamnophilidae): the *Hypocnemis cantator* complex. *Auk* **124**: 11–28.

Isler, M.L., Isler, P.R., Whitney, B.M., Zimmer, K.J. & Whittaker, A. 2009. Species limits in antbirds (Aves: Passeriformes: Thamnophilidae): an evaluation of *Frederickena unduligera* (Undulated Antshrike) based on vocalizations. *Zootaxa* **2305**: 61–68.

IUCN Species Survival Commission. 2001. *IUCN Red List Categories and Criteria: Version 3.1*. Gland, Switzerland and Cambridge, UK: IUCN (http://www.iucnredlist.org/static/categories_criteria_3_1).

Johnson, N.K., Remsen, J.V. & Cicero, C. 1999. Resolution of the debate over species concepts in ornithology: a new comprehensive biologic species concept. *Proc. Int. Ornithol. Congr* **22**: 1470–1482.

Joseph, L. & Omland, K.E. 2009. Phylogeography: its development and impact in Australo-Papuan ornithology with special reference to paraphyly in Australian birds. *Emu* **109**: 1–23.

Knowles, L.L. & Carstens, B.C. 2007. Delimiting species without monophyletic gene trees. *Syst. Biol.* **56**: 887–895.

Konishi, M. 1970. Evolution of design features in the coding of species specificity. *Am. Zool.* **10**: 67–72.

Leaché, A.D., Koo, M.S., Spencer, C.L., Papenfuss, T.J., Fisher, R.N. & McGuire, J.A. 2009. Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (*Phrynosoma*). *Proc. Natl Acad. Sci. USA* **106**: 12418–12423.

Lee, M.S.Y. 2003. Species concepts and species realities: salvaging a Linnaean rank. *J. Evol. Biol.* **15**: 179–188.

Li Shou-Hsien, L., Li Jing-Wen, L., Han Lian-Xian, H., Yao Cheng-Te, Y., Shi Haitao, S., Lei, F.-M. & Yen, C. 2006. Species delimitation in the Hwamei *Garrulax canorus*. *Ibis* **148**: 698–706.

Losos, J.B., Warheit, K.I. & Schoener, T.W. 1997. Adaptive differentiation following experimental island colonisation in *Anolis* lizards. *Nature* **387**: 70–73.

Mace, G.M. 2004. The role of taxonomy in species conservation. *Philos. Trans. R. Soc. Lond. B* **359**: 711–719.

Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends Ecol. Evol.* **20**: 229–237.

Mallet, J. 2008. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philos. Trans. R. Soc. Lond. B* **363**: 2971–2986.

Marks, B.B., Hackett, S.J. & Capparella, A.P. 2002. Historical relationships among Neotropical lowland forest areas of endemism as determined by mitochondrial DNA sequence variation within the Wedge-billed Woodcreeper (Aves: Dendrocolaptidae: *Glyphorynchus spirurus*). *Mol. Phylogenet. Evol.* **24**: 153–167.

Marler, P. 1957. Specific distinctiveness in communication signals of birds. *Behaviour* **11**: 13–39.

Marler, P. & Slabbekoorn, H. (**eds**) 2004. *Nature's Music: The Science of Birdsong*. London: Academic Press.

Martens, J., Eck, S., Päckert, M. & Sun, Y.-H. 2003. Methods of systematic and taxonomic research on passerine birds: the timely example of the *Seicercus burkii* complex (Sylviidae). *Bonn. Zool. Beitr.* **51**: 109–118.

May, R.M. 1988. How many species are there on Earth? *Science* **247**: 1441–1449.

May, R.M. 1990. Taxonomy as destiny. *Nature* **347**: 129–130.

Mayr, E. 1942. *Systematics and the Origin of Species*. New York: Columbia University Press.

Mayr, E. 1969. *Principles of Systematic Zoology*. New York: McGraw-Hill.

Mayr, E. & Gilliard, E.T. 1952. Altitudinal hybridization in New Guinea honeyeaters. *Condor* **54**: 325–337.

McCarthy, E.M. 2006. *Handbook of Avian Hybrids*. Oxford: Oxford University Press.

McCormack, J.E. & Smith, T.B. 2008. Niche expansion leads to small-scale adaptive divergence along an elevation gradient in a medium-sized passerine bird. *Proc. R. Soc. Lond. B* **275**: 2155–2164.

McKay, B.D. & Zink, R.M. 2010. The causes of mitochondrial gene tree paraphyly in birds. *Mol. Phylogenet. Evol.* **54**: 647–650.

Meiri, S. & Mace, G.M. 2007. New taxonomy and the origin of species. *PLoS Biol.* **5**: 1385–1386.

Meyer, C.P. & Paulay, G. 2005. DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol.* **3**: 2229–2238.

Miller, M.J., Bermingham, E., Klicka, J., Escalante, P., do Amaral, F.S., Weir, J.T. & Winker, K. 2008. Out of Amazonia again and again: episodic crossing of the Andes promotes diversification in a lowland forest flycatcher. *Proc. R. Soc. Lond. B* **275**: 1133–1142.

Mishler, B.D. & Donoghue, M.J. 1982. Species concepts: a case for pluralism. *Syst. Zool.* **31**: 491–503.

Moritz, C. 1994. Defining 'Evolutionary Significant Units' for conservation. *Trends Ecol. Evol.* **9**: 373–375.

Moritz, C. & Cicero, C. 2004. DNA barcoding: promise and pitfalls. *PLoS Biol.* **2**: 1529–1531.

Nakagawa, S. & Cuthill, I.C. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* **82**: 591–605.

Nordal, I. & Stedje, B. 2005. Paraphyletic taxa should be accepted. *Taxon* **54**: 5–6.

Nyári, Á.S. 2007. Phylogeographic patterns, molecular and vocal differentiation, and species limits in *Schiffornis turdina* (Aves). *Mol. Phylogenet. Evol.* **44**: 154–164.

Omland, K.E., Tarr, C.L., Boarman, W.I., Marzluff, J.M. & Fleischer, R.C. 2000. Cryptic genetic variation and paraphyly in ravens. *Proc. R. Soc. Lond. B* **267**: 2475–2482.

Ottosson, U., Bensch, S., Olsson, U., Svensson, L. & Waldenström, J. 2005. Differentiation and phylogeny of the Olivaceous Warbler *Hippolais pallida* species complex. *J. Ornithol.* **146**: 127–136.

Päckert, M., Martens, J., Kosuch, J., Nazarenko, A. & Veith, M. 2003. Phylogenetic signal in the song of crests and kinglets (Aves: *Regulus*). *Evolution* **57**: 616–629.

Patten, M.A. & Unitt, P. 2002. Diagnosability versus mean differences of Sage Sparrow subspecies. *Auk* **119**: 26–35.

Patten, M.A., Rotenberry, J.T. & Zuk, M. 2004. Habitat selection, acoustic adaptation, and the evolution of reproductive isolation. *Evolution* **58**: 2144–2155.

Pavlova, A., Zink, R.M., Drovetski, S.G., Red'kin, Y. & Rohwer, S. 2003. Phylogeographic patterns in *Motacilla flava* and *Motacilla citreola*: species limits and population history. *Auk* **123**: 744–758.

Payne, R.B. & Sorenson, M.D. 2007. Integrative systematics at the species level: plumage, songs and molecular phylogeny of Quail-finch *Ortygospiza*. *Bull. Br. Orn. Club* **127**: 4–26.

Peterson, A.T. 1998. New species and new species limits in birds. *Auk* **115**: 555–558.

Peterson, A.T. & Moyle, R.G. 2008. An appraisal of recent taxonomic reappraisals based on character scoring systems. *Forktail* **24**: 110–112.

Phillimore, A.B. & Owens, I.P.F. 2006. Are subspecies useful in evolutionary and conservation biology? *Proc. R. Soc. Lond. B* **273**: 1049–1053.

Phillimore, A.B., Owens, I.P.F., Black, R.A., Chittock, J., Burke, T. & Clegg, S.M. 2008. Complex patterns of genetic and phenotypic divergence in an island bird and the consequences for delimiting conservation units. *Mol. Ecol.* **17**: 2839–2853.

Price, T.D. 2008. *Speciation in Birds*. Greenwood Village, CO: Roberts & Company.

Price, T.D. & Bouvier, M.M. 2002. The evolution of F1 postzygotic incompatibilities in birds. *Evolution* **56**: 2083–2089.

de Queiroz, K. 1998. The general lineage concept of species, species criteria, and the process of speciation. In Howard, D.J. & Berlocher, S.H. (eds) *Species, New Interdisciplinary Essays*: 57–75. Oxford: Oxford University Press.

de Queiroz, K. 2005. Ernst Mayr and the modern concept of species. *Proc. Natl Acad. Sci. USA* **102**: 6600–6607.

de Queiroz, K. 2007. Species concepts and species delimitation. *Syst. Biol.* **56**: 879–886.

Remsen, J.V. 2005. Pattern, process and rigor meet classification. *Auk* **122**: 403–413.

Rheindt, F.E., Norman, J.A. & Christidis, L. 2008. DNA evidence shows vocalizations to be a better indicator of taxonomic limits than plumage patterns in *Zimmerius* Tyrant-flycatchers. *Mol. Phylogenet. Evol.* **48**: 150–156.

Rissler, L.J. & Apodaca, J.J. 2007. Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*). *Syst. Biol.* **56**: 924–942.

Rodrigues, A.S.L., Pilgrim, J.D., Lamoreux, J.F., Hoffmann, M. & Brooks, T.M. 2006. The value of the IUCN Red List for conservation. *Trends Ecol. Evol.* **21**: 71–76.

Rolshausen, G., Segelbacher, G., Hobson, K.A. & Schaefer, H.M. 2009. Contemporary evolution of reproductive isolation and phenotypic divergence in sympatry along a migratory divide. *Curr. Biol.* **19**: 2097–2101.

Rosen, D.E. 1978. Vicariant patterns and historical explanation in biogeography. *Syst. Zool.* **27**: 159–188.

Rosenberg, N.A. 2007. Statistical tests for taxonomic distinctiveness from observations of monophyly. *Evolution* **61**: 317–323.

Rowlett, J. 2003. [Review of] 'The Birds of Ecuador, Volume I: Status, Distribution, and Taxonomy. Volume II: Field Guide'. *Auk* **120**: 562–568.

Ryder, O.A. 1986. Species conservation and systematics: the dilemma of subspecies. *Trends Ecol. Evol.* **1**: 9–10.

Salzburger, W., Martens, J. & Sturmbauer, C. 2002. Paraphyly of the Blue Tit (*Parus caeruleus*) suggested from cytochrome b sequences. *Mol. Phylogenet. Evol.* **24**: 19–25.

Sangster, G. 2000a. Genetic distance as a test of species boundaries in the Citril Finch *Serinus citrinella*: a critique and taxonomic reinterpretation. *Ibis* **142**: 487–490.

Sangster, G. 2000b. Taxonomic stability and avian extinctions. *Conserv. Biol.* **14**: 579–581.

Schluter, D. 2009. Evidence for ecological speciation and its alternative. *Science* **323**: 737–741.

Seddon, N. & Tobias, J.A. 2010. Character displacement from the receiver's perspective: species and mate-recognition despite convergent signals in suboscine birds. *Proc. R. Soc. Lond. B* **277**: 2475–2483.

Seddon, N., Tobias, J.A., Eaton, M. & Ödeen, A. 2010. Human vision can provide a valid proxy for avian perception of sexual dichromatism. *Auk* **127**: 283–292.

Shaffer, H.B. & Thomson, R.C. 2007. Delimiting species in recent radiations. *Syst. Biol.* **56**: 896–906.

Sites, J.W. & Marshall, J.C. 2003. Delimiting species: a Renaissance issue in systematic biology. *Trends Ecol. Evol.* **18**: 462–470.

Sites, J.W. & Marshall, J.C. 2004. Operational criteria for delimiting species. *Ann. Rev. Ecol. Evol. Syst.* **35**: 199–227.

Slabbekoorn, H. & Smith, T.B. 2002. Bird song, ecology and speciation. *Philos. Trans. R. Soc. Lond. B* **357**: 493–503.

Smith, A.L., Monteiro, L., Hasegawa, O. & Friesen, V.L. 2007. Global phylogeography of the Band-rumped Storm-petrel (*Oceanodroma castro*; Procellariiformes: Hydrobatidae). *Mol. Phylogenet. Evol.* **43**: 755–773.

Sorenson, M.D., Sefc, K.M. & Payne, R.B. 2003. Speciation by host switch in brood parasitic Indigobirds. *Nature* **424**: 928–931.

Stebbins, G.L. 1969. Comments on the search for a 'perfect system'. *Taxon* **18**: 357–359.

Talbot, S.L. & Shields, G.F. 1996. Phylogeography of brown bears (*Ursus arctos*) of Alaska and paraphyly within the Ursidae. *Mol. Phylogenet. Evol.* **5**: 477–494.

Techow, N.M.S.M., Ryan, P.G. & O'Ryan, C. 2009. Phylogeography and taxonomy of White-chinned and Spectacled Petrels. *Mol. Phylogenet. Evol.* **52**: 25–33.

Theron, E., Hawkins, K., Bermingham, E., Ricklefs, R.E. & Mundy, N.I. 2001. The molecular basis of an avian plumage polymorphism in the wild: a point mutation in the melanocortin-1 receptor is perfectly associated with melanism in the Bananaquit (*Coereba flaveola*). *Curr. Biol.* **11**: 550–557.

Tobias, J.A. & Seddon, N. 2009. Signal design and perception in *Hypocnemis* antbirds: evidence for convergent evolution via social selection. *Evolution* **63**: 3168–3189.

Tobias, J.A., Bates, J.M., Hackett, S.J. & Seddon, N. 2008. Comment on the latitudinal gradient in recent speciation and extinction rates of birds and mammals. *Science* **319**: 901.

Toews, D.P.L. & Irwin, D.E. 2008. Cryptic speciation in a Holarctic passerine revealed by genetic and bioacoustic analyses. *Mol. Ecol.* **17**: 2691–2705.

Wang, J.Y., Frasier, T.R., Yang, S.C. & White, B.N. 2008. Detecting recent speciation events: the case of the finless porpoise (genus *Neophocaena*). *Heredity* **101**: 145–155.

Wiens, J.J. 2007. Species delimitation: new approaches for discovering diversity. *Syst. Biol.* **56**: 875–878.

Wiens, J.J. & Servedio, M.R. 2000. Species delimitation in systematics: inferring diagnostic differences between species. *Proc. R. Soc. Lond. B* **267**: 631–636.

Will, K.W., Mishler, B.D. & Wheeler, Q.D. 2005. The perils of DNA barcoding and the need for integrative taxonomy. *Syst. Biol.* **54**: 844–851.

Winker, K. 2009. Reuniting genotype and phenotype in biodiversity research. *Bioscience* **59**: 657–665.

Winker, K. 2010. Subspecies represent geographically partitioned variation, a goldmine of evolutionary biology, and a challenge for conservation. *Ornithol. Monogr.* **67**: 6–23.

Winker, K., Rocque, D.A., Braile, T.M. & Pruett, C.L. 2007. Vainly beating the air: species-concept debates need not impede progress in science or conservation. *Ornithol. Monogr.* **63**: 30–44.

Yoder, A.D., Olson, L.E., Hanley, C., Heckman, K.L., Rasoloarison, R., Russell, A.L., Ranivo, J., Soarimalala, V., Karanth, K.P., Raselimanana, A.P. & Goodman, S.M. 2005. A multidimensional approach for detecting species patterns in Malagasy vertebrates. *Proc. Natl Acad. Sci. USA* **102**: 6587–6594.

Zander, R.H. 2007. Paraphyly and the species concept, a reply to Ebach & al. *Taxon* **56**: 642–644.

Zink, R.M. 1994. The geography of mitochondrial DNA variation, population structure, hybridization and species limits in the Fox Sparrow (*Passerella iliaca*). *Evolution* **48**: 96–111.

Zink, R.M. 2006. Rigor and species concepts. *Auk* **123**: 887–891.

Zink, R.M. & McKitrick, M.C. 1995. The debate over species concepts and its implications for ornithology. *Auk* **112**: 701–719.

Zink, R.M., Pavlova, A., Drovetski, S., Wink, M. & Rohwer, S. 2009. Taxonomic status and evolutionary history of the *Saxicola torquata* complex. *Mol. Phylogenet. Evol.* **52**: 769–773.

Received 14 September 2009;
revision accepted 13 July 2010.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table S1. A tabulated summary of the British Ornithologists' Union guidelines for assigning species rank in birds.

Appendix S1. A detailed critique of the British Ornithologists' Union guidelines for assigning species rank in birds.

Appendix S2. Full dataset of biometric, vocal and plumage information for all 58 species and 24 subspecies used in analyses.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.