J. Field Ornithol. 87(1):104–115, 2016 DOI: 10.1111/jofo.12141

COMMENTARY

Toward a scoring system for species delimitation: a response to Remsen

Nigel J. Collar,[1] Lincoln D. C. Fishpool,[1] Josep del Hoyo,[2] John D. Pilgrim,[3] Nathalie Seddon,[4] Claire N. Spottiswoode,[5] and Joseph A. Tobias[6,7]

[1] *BirdLife International, Girton Road, Cambridge CB3 0NA, United Kingdom*
[2] *Lynx Edicions, Montseny 8, E-08193 Bellaterra, Spain*
[3] *The Biodiversity Consultancy, 3E King's Parade, Cambridge CB2 1SJ, United Kingdom*
[4] *Department of Zoology, Oxford University, Oxford OX1 3PS, United Kingdom*
[5] *Department of Zoology, Cambridge University, Cambridge CB2 3EJ, United Kingdom*
[6] *Department of Life Sciences, Imperial College London, Silwood Park, Buckhurst Road, Ascot, Berkshire SL5 7PY, United Kingdom*

A recent review by Remsen (2015) of the HBW–BirdLife International illustrated checklist of the birds of the world (hereafter, the Checklist) contained some very positive comments about the book itself, but some very negative ones about the system for species delimitation that it employs. Although it is normal and appropriate for authors and publishers of books to not take issue with their reviewers, in this case, some 80% of Remsen's review involved a critique of another study that was published in a peer-reviewed journal 5 yr before (Tobias et al. 2010). It is this work, outlining what the Checklist calls "the Tobias criteria," that we here primarily seek to defend. However, because Remsen (2015) sometimes implies criticism of the authors of the Checklist as well as, or instead of, the work reported in Tobias et al. (2010), we also take the opportunity to reflect on some of these points.

To recapitulate, the Tobias criteria were developed after calibrating the degree of phenotypic difference shown by 58 pairs of species (from all continents and latitudes) identified by experts as examples of lineages that are both phenotypically similar and each other's closest relatives in sympatry (and, therefore, not necessarily sister species). The criteria allow scores of 1–4 to reflect the increasing strength of characters considered, and restrict scoring to three plumage, two morphometric, and two acoustic characters, with scores up to 2 for ecological and behavioral differences. In addition, scores of 1, 2, and 3 are given for, respectively, broad hybrid zone, narrow hybrid zone, and line of parapatry between taxa. Scores are weighted to emphasize the importance of traits functioning as social or mating signals. A score of seven points elevates a subspecies to species status. When published, this system was heralded as an "operational revolution in taxonomy" (Brooks and Helgen 2010), and as firmly placing "a degree of consistency and transparency upon taxonomic decisions" (Winker 2010), with its use of effect sizes representing "a crucial step in the direction of objectivity" (Patten 2015).

Remsen (2015) nevertheless rejects the taxonomic revisions proposed in the Checklist, all of which are based on this system, but reaches this position on the back of some unfortunate misconceptions. (1) "The Tobias... scheme requires making decisions based on comparisons to close relatives" (p. 185). This is diametrically opposite to the truth. Comparisons with close relatives are used by other systems, including Mayr et al. (1953) and Helbig et al. (2002), and are sometimes informally made by molecular scientists proposing changes in taxonomic rank, but are often problematic because there are either too many options or none at all. The seven-point system avoids this difficulty by requiring no comparison other than between the taxa being scored (Tobias et al. 2010: 740). (2) "The Tobias... scheme emphasizes the importance of adjusting their scheme for the group involved"

[7] Corresponding author. Email: j.tobias@imperial.ac.uk

(p. 186). This is also a mistake; an entire paragraph in Tobias et al. (2010: 740) is dedicated to explaining why no such adjustment is needed or even desirable. Remsen (2015) states that "Although not specifically stated, apparently the idea is that whatever *Empidonax* lack in plumage characters is counterbalanced by diversity in vocal characters"—yet this *is* specifically, precisely stated in Tobias et al. (2010), even to the point of invoking the case of *Empidonax* flycatchers, that is, "divergence at the species level is fairly consistent when the full array of traits is taken into account" (empirical data are presented to support this case). (3) "The authors . . . assign species to narrow or broad [hybrid] zones . . . without any justification, much less biological rationale, for such an utterly arbitrary scheme" (p. 187). This overlooks the passage in Tobias et al. (2010: 731) that explains the 200-km threshold for such assignments is derived from Price's (2008) analysis of 23 hybrid zones (mean width = 224 km). The threshold may be arbitrary, and is certainly debatable, but it is emphatically not without "justification [or] biological rationale."

There is a serious fallacy, too, underlying Remsen's (2015) suspicion over the way the Tobias criteria were validated. He worries that the selection of "23 European taxa ranked as subspecies" for this exercise might represent a "biased sample" from which "extrapolations could be perilous." However, the validity of the exercise depended entirely on the use of a taxonomically stable avifauna, to check that the seven-point threshold gave results that conform to a well-established and widely accepted norm in the application of the Biological Species Concept, and it was, indeed, notable that the only two taxa to emerge as species were not European, but North African, so there is no basis for the criticism that the "knowns" of the European avifauna represent a "biased sample." Rather, they conform to the threshold established by a much larger sample of taxa from all continents, including tropical latitudes. Most importantly, no extrapolations were made from western Palearctic birds, that is, they were not used to set or adjust the threshold, but simply as a test case to validate the approach. Indeed, one of the stated aims of the criteria is to provide a benchmark so that the standards used in the well-studied temperate zone are applied to taxonomic revisions worldwide. There is thus no peril and no justification for suggesting otherwise.

Remsen (2015) then asks why the criteria could not boost the threshold to "eliminate the outliers . . . to produce a more conservative approach," but this question implies unswerving loyalty to the *status quo* for no clear reason. Our method aligned with 95% of accepted taxonomic decisions in the western Palearctic and conflicted with 5% of cases. There is a logical basis for accepting an error rate of 5%, this being the standard critical value for statistical significance in most biological studies. Conversely, Remsen's (2015) suggestion that we adjust our thresholds to the very tip of the tail of the distribution seems to rest on faulty logic because those outliers will be subject to the vagaries of taxonomic error and uncertainty. In light of this, we fail to see the virtue in "a more conservative approach," which sounds suspiciously like a value judgment rather than an expression of scientific objectivity. A score well above 7—an outlier—would represent evidence of a taxon's substantial distinctiveness, so why would anyone want to obscure it by designating a higher threshold?

Remsen (2015) introduces the next step in his critique by considering three cases to "try to understand the mechanics of the system." First up is *Larus smithsonianus* (American Herring Gull), which is more about the Checklist's decision not to apply the criteria than the criteria themselves. Remsen (2015) is happy to cite quotations that indicate the split of this taxon is tricky because of (1) the use of mtDNA and (2) its imperceptible morphological distinctiveness (in adult plumage). However, he fails to credit the Checklist for its frank indication that the taxonomy of Herring Gulls represents "one of the most complex challenges in systematic ornithology," or that its acceptance of *smithsonianus* is provisional, that is, "Molecular work places present form at a distance from other taxa in the complex [four references], and current trend to accept this arrangement followed here pending further investigations." Moreover, in the Checklist Introduction (p. 38), the specific case of *smithsonianus* is discussed:

> The gull is recognized entirely owing to several molecular studies rendering it, somewhat surprisingly, paraphyletic with the near-identical European Herring Gull *L. argentatus*, and with the caveat that this is, owing to the very small genetic distances involved in most of the "Herring Gull complex", not a very satisfactory (and probably not a very stable) arrangement.

Here, then, is a fully acknowledged instance where molecular evidence is tentatively accepted over the likely result (which apparently Remsen would have supported) of an application of the Tobias criteria. Although we hesitate to follow Remsen in describing such evidence as "controversial if not flawed," the explicit caution over the robustness of the arrangement adopted in this case ought surely to have prevented any open-minded reader from finding the decision to recognize *smithsonianus* to be "perplexing." Given Remsen's critique, it is rather more perplexing to discover that *L. smithsonianus* is recognized as a species, comprising exactly the same three subspecies as in the Checklist, in Dickinson and Remsen (2013).

Concerning the Northern (*Patagioenas fasciata*) and Southern (*P. albilinea*) Band-tailed Pigeons, Remsen (2015) suggests that the Tobias criteria are misapplied in the Checklist to produce a score of 8 for this split; he believes the score should be 7 because of the cap of 3 on the scorable number of plumage and bare part characters (see the second paragraph above). He is both right and wrong here; the score given is 7, but, in the application, all character distinctions are listed and assessed for completeness, with redundant differences taking the prefix (ns) for "no score," as is indicated in the worked example in the Checklist Introduction. Remsen (2015) also criticizes our emphasis on plumage pattern in this case, on the grounds that voice is more important in New World pigeons. This may or may not be true, but it misses the point of the Tobias criteria. If a taxon scores 7 on plumage/morphometric characters, as in this case, the vocal evidence is simply not needed for its elevation to species rank; conversely, if it scores 7 on voice, evidence is then not needed from plumage/morphometrics. There is no provision—nor should there be—for a downward revision of a score just because vocal traits between taxa turn out to be similar. Remsen (2015) then proposes that Northern Band-tailed Pigeons can sing like Southern Band-tailed Pigeons, citing as evidence a publicly available recording (Macaulay Library 70842). However, this sound file—originally recorded by Ted Parker, but identified subsequently by someone else—was made outside the geographical and elevational range of any "Band-tailed Pigeon," and apparently contains the song of a Grey-headed Dove (*Leptotila plumbeiceps*). Meanwhile, acoustic analysis of 10 recordings of the two main races of *P. fasciata* and 12 recordings of all three races of *P. albilinea* reveals consistent vocal differences, producing a score of 3 for significant disparity in maximum frequency and 3 for bisyllabic versus monosyllabic call (Boesman 2015), thereby strongly vindicating the morphological evidence. (We note that revisions in the Checklist non-passerine volume often focused on plumage divergence for the sake of expediency; the emphasis on quantitative assessments of vocal divergence will be stronger in the passerine volume in preparation.)

The issue of hybridization flows from this point through the remainder of Remsen's (2015) review in various guises, including the sections "The signal from contact zones is ignored" and "Non-assortative and assortative mating are not distinguished." We note that frequent hybridization can occur between perfectly valid species, as now enshrined in modern versions of the Biological Species Concept (see Johnson et al. 1999). However, the question of how much hybridization is allowable before species are demoted to subspecies is open to debate. Remsen's (2015) view appears to be that truly non-assortative mating in contact zones is indicative of subspecies status because the two taxa do not treat each other as species (so, "why should we?"). This argument sounds persuasive, but is it watertight?

There is no doubt that high levels of hybridization between lineages indicate that premating reproductive isolation is absent or very incomplete, but such lineages may nonetheless qualify as species if their evolutionary independence is maintained by post-mating reproductive isolation. This is clarified by studies of taxa such as Lazuli and Indigo buntings (*Passerina amoena* and *P. cyanea*), and Pied and Collared flycatchers (*Ficedula hypoleuca* and *F. albicollis*) that are universally accepted as species, but often hybridize in their contact zones. In these cases, behavioral interactions and rates of interbreeding tell us less about species status than do the genomic incompatibilities maintaining the hybrid zone (Carling and Brumfield 2008, Harr and Price 2012). We would, therefore, argue that evolutionary lineages may be species even when they do not treat each other as such. Even introgression of genes and associated phenotypes outside the contact zone is arguably of dubious importance because introgression may be limited to compatible genes, whereas the zone is maintained by non-introgression of incompatible genes (Endler 1977).

The case of Red-shafted/Yellow-shafted flickers (*Colaptes*), and other examples of distinctive lineages hybridizing non-assortatively in contact zones (e.g., *Ramphastos* and *Pionites* in Amazonia), will always be contentious. We agree with Remsen that non-assortative mating implies that plumage signals are not operating effectively as reproductive isolating mechanisms. We also acknowledge that hybrid viability is doubtless high, at least locally, in these cases. However, as pointed out in the Checklist Introduction (p. 33), "If hybrids were fully viable, genomes fully compatible and signals not reproductively isolating, then the contact zone between two hybridizing taxa would be a broad cline." Our approach is thus based on the assumption that a restricted hybrid zone, by its very existence, indicates that some combination of pre- and post-mating isolation must exist to maintain the evolutionary independence of core lineages. If such lineages are phenotypically distinctive, the chance of them persisting through time is high, and we thus prefer to classify them as separate species despite non-assortative mating (Helbig et al. 2002). This approach simply extends the standard taxonomic treatment applied to other distinctive lineages, for example, Yellowhammer (*Emberiza citrinella*) and Pine Bunting (*E. leucocephalos*), which are classified as species despite interbreeding freely in relatively broad contact zones (Copete 2011).

Remsen (2015) also argued that parapatry is "sufficient evidence of species rank," and thus that our scoring of parapatry makes no sense. However, the theory and reality of parapatric distributions are rather different. In theory, a line of parapatry is as narrow as that separating the adjacent territories of two mutually exclusive taxa, in which case the taxa must be species. In reality, it is never so simple. Remsen might have had an easier time "taking the rest of the scheme seriously" if he had consulted the online supplementary material (Appendix S1) in Tobias et al. (2010), where an entire section is devoted to the complexities of parapatry, including a rationale for its treatment in the scoring system. Among other things, we point out that viewing parapatry as contact "without free gene flow" is overly simplistic:

> The impression of abutting ranges may simply relate to inadequate sampling when boundaries are defined by point localities. Problems may also arise when apparent boundaries coincide with physical features: the BOU guidelines [Helbig et al. 2002] refer to rivers as "trivial barriers", but this is not necessarily true for tropical forest passerines, some of which are unwilling or indeed unable to cross small stretches of water (Moore et al. 2008).

Are there, in fact, any cases in the world where we can say with certainty that parapatry is the condition, rather than a very narrow hybrid zone or a very narrow line with no contact? The subtly, but indisputably distinct Icterine and Melodious warblers (*Hippolais icterina* and *H. polyglotta*) replace each other geographically, but form a narrow hybrid zone in the process (Bairlein 2006). How would the relationship even be noticed in poorly known regions and for adjacent taxa that are still more look-alike and sound-alike? In addition, what would be the chances that they never interbreed? Conversely, if they are really so indistinct *and* they form a hybrid zone, would this not indicate that they merit subspecies rank only—which is what the Tobias system would tell us? Remsen's (2015) argument that "the scoring process is backwards" assumes that we know what happens in cases of apparent parapatry. In reality, we often do not, so itemizing the level of difference between putatively parapatric taxa is an important component in a dependable and responsible taxonomic appraisal.

The apex of Remsen's (2015) critique is his headline assertion that "Phenetic taxonomy has risen from the dead," but did it ever die? The closely related fields of phenetics and numerical taxonomy rely on quantitative methods based on character similarity, and have largely been replaced by cladistics as a method to estimate relationships among taxa, often with phylogenetic techniques. This is perfectly understandable because phenetics struggles to deal with widespread evolutionary phenomena such as symplesiomorphies (the recurrence of ancestral traits in distantly related lineages or clades), which can render character comparisons phylogenetically uninformative (Futuyma 1998). However, far from being "long-abandoned," the use of phenetic taxonomy to assess species limits (taxonomic ranks) between pairs of related lineages remains widespread because, in these instances, the issue of symplesiomorphy is irrelevant. Numerous species and subspecies have been described and accepted in recent decades based exclusively on levels of phenetic differentiation. Moreover, the scoring system employed in the Checklist bears little relation to the phenetic techniques that Remsen disdains because it is weighted toward traits likely to play a role in maintaining reproductive isolation, and the number and type of characters included are specified and capped to avoid the inflation and bias of scores.

For these reasons, Remsen's comments about phenetics are distinctly off-key, and undermine his conclusion that the Tobias system is "conceptually flawed" because "a tally of the number of character differences . . . determines taxon rank." Ironically, tallying numbers is precisely the approach adopted in the well-established system outlined by Isler et al. (1998), which has been widely applied to the determination of taxon rank in antbirds and that advises that "a focus on the *number of characters* [our italics] is appropriate given the possibility that the role and importance of vocalization types in species' repertoires may differ across groups of taxa." Remsen (2005) actually published a commentary about Isler et al. (1998), commending its "rigor" and concluding that "this methodology allows an objective, reproducible classification of allopatric thamnophilid taxa." Moreover, all the resulting species-level determinations using that system have been accepted by the South American Checklist Committee (SACC), of which Remsen is the chair.

Other Remsen (2015) criticisms include the citation of unpublished Ph.D. dissertations (we counter that such material has undergone scientific scrutiny and is in the public domain so a responsibility exists to make use of it, where informative), and the implication that the Checklist authors "cherry-picked" the results of DNA studies (no instance of this is given, however). Similarly, the remark (p. 183) that "the authors restricted their analyses to a subset of the world's avifauna chosen by their own expertise" carries an unwarranted implication of bias; it is true that not all potential cases were examined, but every effort was made to assess those where the available evidence was indicative, independent of our own "expertise." Remsen appears to accept the authors' assertion that no conservation considerations biased their taxonomic assessments, but a reviewer pondering this issue might have checked for evidence and noted, for example, resistance to the elevation of some parrot lineages (e.g., *Poicephalus robustus robustus* and *Pezoporus wallicus flaviventris*) to species status, despite these cases being strongly argued by conservation-conscious authorities (e.g., Perrin 2005, Murphy et al. 2011).

The cumulative effect of Remsen's (2015) objections and reservations leads him to recommend "blanket dismissal of the novel rank assignments in this volume." However, as this examination of them shows, most of his concerns are invalid and none is fatal. We, therefore, fail to see the objective basis for such a sweeping exhortation. Moreover, we note that Remsen has independently assessed at least two of the "rank assignments" made by the Checklist and, in both cases, endorsed them. First, reviewing the four-way split of the *Oxypogon* helmetcrests based on the seven-point system (Collar and Salaman 2013), Remsen remarked "I can't think of any other genus in which males differ to this degree yet are ranked as subspecies" (Remsen et al. 2015). Second, the Checklist split the Purple-/Violet-crowned Plovercrest (*Stephanoxis loddigesii*) from Green-crowned Plovercrest (*S. lalandii*), with a rather high score of 11 (an outlier indeed!), simultaneously with and in ignorance of Cavarzere et al. (2014). On the basis of Cavarzere et al.'s (2014) evidence, but noting the concordant treatment independently given in the Checklist, Remsen proposed this split to SACC with the comment "In my opinion, by any reasonable standard of comparative degree of phenotypic divergence for allotaxa, these should be treated as separate species." Given his agreement in these cases, how then can he decide that the several hundred original taxonomic changes presented in the Checklist are not worth serious examination?

These examples highlight a key point: the scoring system used in the Checklist commonly reaches conclusions consistent with other perceptions on species limits. As painstakingly explained in Tobias et al. (2010), the method is not proposed as a flawless technique for accurate species delimitation. It is a relatively simple rule-of-thumb for accelerating taxonomic revisions, allowing us to get a firmer and more consistent grip on species diversity, particularly in regions (for example, the Asian tropics) where levels of taxonomic activity are relatively modest (Collar

2003). Consequently, the scoring system is most appropriate for cases where certain relevant data are unavailable, and the fact that it is easily overruled by detailed research or phylogenetic studies is an integral part of the method. If, for example, reliable studies confirm that populations are in contact, and genetic data indicate that introgression is minimal, then two species are clearly involved and the scoring system is redundant.

The Checklist non-passerine volume contains a large number of taxonomic revisions, with their justifications. These result from a considerable investment of time and effort to accumulate the evidence. Nothing in Remsen's critique of the Tobias criteria invalidates the judgments that have been placed on that evidence and, consequently, we urge that, as is normal practice in science, the revisions presented in the Checklist be submitted to independent case-by-case review to evaluate them on their merit. This is, we believe, by far the more objective and rational approach to the situation. The appropriate scientific response must always be to consider the evidence, not willfully ignore it.

Remsen (2005) once mused that "recent attention to conservation of biodiversity could catalyze quantitative, comprehensive overhauls of subspecies taxonomy." Indeed, the seven-point system was partly inspired by his rallying cry that "alternatives are needed" to speed up the taxonomic process. Our quantitative approach provides a system that does not claim to be flawless and which welcomes constructive criticism; numerous sources of potential subjectivity and possibilities for improvement are discussed in Tobias et al. (2010). Meanwhile, however, we suspect that accepted species limits in coming decades will be rather different from current taxonomic treatments, and far closer to the revisions proposed in the Checklist. Fast-tracking these taxonomic decisions, evaluating their accuracy, and fine-tuning the process by which they are made should be priorities for museum scientists, conservationists, policy makers, and the community of field ornithologists this journal serves.

LITERATURE CITED

BAIRLEIN, F. 2006. Family Sylviidae (Old World warblers). In: Handbook of the birds of the world, vol. 11. Old World Flycatchers to Old World Warblers (J. del Hoyo, A. Elliott, and D. A. Christie, eds.), pp. 492–709. Lynx Edicions, Barcelona, Spain.

BOESMAN, P. [online]. 2015. Voice comparison of the Southern and Northern Band-tailed Pigeons. HBW Alive Ornithological Note 41. In: Handbook of the birds of the world alive. Lynx Edicions, Barcelona, Spain. <http://www.hbw.com/node/909144> (Accessed 20 August 2015).

BROOKS, T. M., AND K. M. HELGEN. 2010. A standard for species. Nature 467: 540–541.

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

CAVARZERE, V., L. F. SILVEIRA, M. F. DE VASCONCELOS, R. GRANTSAU, AND F. C. STRAUBE. 2014. Taxonomy and biogeography of *Stephanoxis* Simon, 1897 (Aves: Trochilidae). Papéis Avulsos de Zoologia 54: 69–79.

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

———, AND P. G. W. SALAMAN. 2013. The taxonomic and conservation status of the *Oxypogon* helmetcrests. Conservación Colombiana 19: 31–38.

COPETE, J. 2011. *Emberiza citrinella* and *E. leucocephalos*. In: Handbook of the birds of the world, vol. 16. Tanagers to New World blackbirds (J. de Hoyo, A. Elliott, and D. A. Christie, eds.), pp. 510–511. Lynx Edicions, Barcelona, Spain.

DICKINSON, E. C., AND J. V. REMSEN, Jr. (eds.). 2013. The Howard & Moore complete checklist of the birds of the world, vol. 1, 4th ed. Aves Press, Eastbourne, UK.

ENDLER, J. A. 1977. Geographic variation, speciation, and clines. Princeton University Press, Princeton, NJ.

FUTUYMA, D. J. 1998. Evolutionary biology, 3rd ed. Sinauer Associates, Sunderland, MA.

HARR, B., AND T. PRICE. 2012. Speciation: clash of the genomes. Current Biology 22: R1044–R1046.

HELBIG, A. J., A. G. KNOX, D. T. PARKIN, G. SANGSTER, AND M. COLLINSON. 2002. Guidelines for assigning species rank. Ibis 144: 518–525.

ISLER, M. L., P. R. ISLER, AND B. M. WHITNEY. 1998. Use of vocalizations to establish species limits in antbirds (Passeriformes: Thamnophilidae). Auk 115: 577–590.

JOHNSON, N. K., J. V. REMSEN, AND C. CICERO. 1999. Resolution of the debate over species concepts in ornithology: a new comprehensive biologic species concept. Proceedings of the International Ornithological Congress 22: 1470–1482.

MAYR, E., E. G. LINSLEY, AND R. L. USINGER. 1953. Methods and principles of systematic zoology, 1st ed. McGraw-Hill, New York, NY.

MOORE, R. P., W. D. ROBINSON, I. J. LOVETTE, AND T. R. ROBINSON. 2008. Experimental evidence for extreme dispersal limitation in tropical forest birds. Ecology Letters 11: 960–968.

MURPHY, S. A., L. JOSEPH, A. H. BURBIDGE, AND J. AUSTIN. 2011. A cryptic and critically endangered species revealed by mitochondrial DNA analyses: the Western Ground Parrot. Conservation Genetics 12: 595–600.

PATTEN, M. A. 2015. Subspecies and the philosophy of science. Auk 132: 481–485.

PERRIN, M. R. 2005. A review of the taxonomic status and biology of the Cape Parrot *Poicephalus robustus*, with reference to the Brown-necked Parrot *P. fuscicollis fuscicollis* and the Grey-headed Parrot *P. f. suahelicus*. Ostrich 76: 195–205.

PRICE, T. D. 2008. Speciation in birds. Roberts & Company, Greenwood Village, CO.

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

———. 2015. [Review of] HBW and BirdLife International Illustrated Checklist of the Birds of the World. Non-passerines (N. J. Collar and J. del Hoyo, eds.), vol. 1, 903 pp. Journal of Field Ornithology 82: 182–187.

REMSEN, J. V., Jr., J. I. ARETA, C. D. CADENA, A. JARAMILLO, M. NORES, J. F. PACHECO, J. PÉREZ-EMÁN, M. B. ROBBINS, F. G. STILES, D. F. STOTZ, AND K. J. ZIMMER [online]. 2015. A classification of the bird species of South America. American Ornithologists' Union. <http://www.museum.lsu.edu/~Remsen/SACCBaseline.html> (Accessed 12 July 2015).

TOBIAS, J. A., N. SEDDON, C. N. SPOTTISWOODE, J. D. PILGRIM, L. D. C. FISHPOOL, AND N. J. COLLAR. 2010. Quantitative criteria for species delimitation. Ibis 152: 724–746.

WINKER, K. 2010. Is it a species? Ibis 152: 679–682.

A "rapid assessment program" for assigning species rank?

J. V. Remsen, Jr.[1,2]

[1]*Museum of Natural Science, Louisiana State University, Baton Rouge, Louisiana 70803, USA*
[2]*Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA*

As noted in my original review (Remsen 2015), I sympathize with the predicament faced by BirdLife International and *Handbook of the Birds of the World*. They are frequently constrained by a species-level taxonomy, especially in the Old World tropics, maintained largely by historical momentum. This outdated taxonomy is often devoid of the perspective of more than 50 years of additional research that shows that minor differences, particularly in vocalizations, are associated with, if not responsible for, barriers to free gene flow. Today, no one would dispute that the broadly defined species of the Peters Check-list era are in dire need of a thorough overhaul to restore much species-level diversity that was "exterminated" in that era by pen strokes without even a phrase of published rationale. As long as conservation decisions are based solely on species rank, those outdated taxonomies have repercussions far beyond avian classification. As is, taxa given little or no conservation attention because they are currently classified as "only" subspecies are potentially in danger of disappearing while waiting for taxonomy to be changed. For those reasons, I appreciate the need for a "rapid assessment program" for revising species limits, for example, Tobias et al. (2010) (hereafter, T10), as implemented by Collar and del Hoyo (2014) (hereafter, CDH14). For the same reasons, however, any implementation of such a program requires careful scrutiny, as in Remsen (2015).

Collar et al. (2016) responded to criticisms by Remsen (2015) largely by taking issue with minor points and interpretation of details. Rather than tediously refute or re-evaluate each of their minor points, I here focus on larger issues largely sidestepped, missed, or obfuscated by Collar et al. (2016).

IGNORING DATA FROM CONTACT ZONES

If the T10 scheme for rapid assessment had been applied by CDH14 to a limited number of data-deficient cases, then my critique would have been much less harsh. However, despite the severe limitations that they themselves pointed out, they applied it universally, including too many cases at the opposite end of the spectrum from "data-deficient." For example, CDH14 applied their ranking scheme to the Red-shafted/Yellow-shafted flicker (*Colaptes*) case, which is perhaps the best-studied contact zone in North America (see Moore 1987, Moore and Price 1993, and references therein). The signal from the flicker contact zone is astoundingly clear, that is, the numerous dramatic plumage differences between the two groups that caused CDH14

Email: najames@lsu.edu

to rank the two flickers as separate species (contra all other current BSC classifications) are irrelevant when it comes to mate choice. Thus, not only is there a hybrid zone 300-km wide, but signs of introgression also extend hundreds of kilometers from the hybrid zone. If CDH14 somehow interpret the empirical data from the flicker hybrid zone as support for *species* rank, then there is really *no point in examining data from any contact zone*. CDH14 dismissed the evidence for rampant gene flow, including support for the Bounded Hybrid Superiority Model (Moore and Koenig 1986), in favor of their scoring scheme for plumage differences between the two pure parental stocks despite strong evidence that these differences are irrelevant to mate choice and most measures of fitness (Moore and Koenig 1986), with consequent blending of the plumage types in the hybrid zone (e.g., Grudzien et al. 1987). Collar et al. (2016) repeat the statement in CDH14 that "If hybrids were fully viable, genomes fully compatible and signals not reproductively isolating, then the contact zone between two hybridizing taxa would be a broad cline." This reveals unawareness of the research on many hybrid zones, including the flickers, which shows that the hybrids are fully viable, have equal or superior fitness in the hybrid zone, and do show clinal distributions of characters of the pure parental types within the hybrid zone. If a hybrid zone is truncated at its boundaries, then this is likely due to decreased fitness beyond those boundaries compared to pure parental genotypes, in the same way that selection presumably maintains discrete genetic and phenotypic core populations at the contact zones between literally hundreds of parapatric taxa that CDH14 rank as subspecies. In other words, if CDH14 treat the flickers as two species, then they should also elevate to species rank hundreds of parapatric non-passerine subspecies that intergrade at their boundaries (and thousands of passerine cases in the forthcoming passerine volume).

This example and the response by Collar et al. (2016) continue to expose the fundamental philosophical difference between their approach to assigning species limits and the approach used by biologists who use measures of gene flow to assess species rank, namely the empirical data from a contact zone should drive any assignment of taxon rank. In contrast, T10 and CDH14 use a phenetic scheme based on phenotypic characters derived from a sample of 58 species of heterogeneous phylogenetic affinities, and then apply these criteria to the pure parental populations between which mate choice is already known or inferred from the distribution of phenotypic characters (and further *add* points in favor of species rank if there is a hybrid zone). When there are no data at all from a contact zone, perhaps then the T10 scheme provides a partly objective, temporary, rapid assessment alternative. Even so, the existing scheme requires major modifications.

PHYLOGENY-FREE COMPARISONS

Collar et al. (2016) continue to applaud independence of phylogeny as a strength of their ranking scheme. That they are impervious to the importance of controlling for phylogenetic effects is reinforced by their response that "an entire paragraph in Tobias et al. (2010: 740) is dedicated to explaining why no such adjustment is needed or even desirable." In contrast, the overwhelming signal from research in evolutionary biology over the last several decades is that "phylogeny matters." Therefore, the use of comparisons by T10 and CDH14 without any control for phylogenetic effects is difficult to defend. (Their defense is that application of their scheme to a sample of 23 pairs of taxa from the Western Palearctic traditionally ranked as subspecies elevates "only" two species to species rank; this falls short in so many obvious ways of representing a validation of their scheme that it is not worth consuming space here to enumerate them.)

The foundation of their comparative framework for assigning pairs of taxa to species or subspecies rank is a set of 58 species pairs listed in T10. However, their sample of 58 congeneric pairs includes only three (5%) from all non-passerine orders combined. In other words, broad application by CDH14 of the T10 criteria to the non-passerine volume was derived from a scoring system that included only three species pairs of non-passerines (one in Musophagiformes and two in Piciformes). Consequently, missing from their yardstick for assigning species rank in non-passerines are, for example, (1) any representatives of any aquatic bird order, (2) any member of the Galloanseres,

(3) any members of any order dominated by nocturnal taxa (where plumage characters are likely of minimal importance), (4) any members of the carnivore orders (Accipitriformes, Strigiformes, and Falconiformes), and (5) any member of species-rich groups such as Psittaciformes, Cuculiformes, Columbiformes, Apodiformes, or Coraciiformes. I think that any contemporary evolutionary biologist would put the burden-of-proof on T10 to show directly that the dramatic phylogenetic skew in their sample has no effect on the outcomes of their analyses. I also predict that it requires no expertise in biology to appreciate the essence of the problem, that is, who cannot appreciate that a scoring scheme derived almost completely from small diurnal passerines might not be appropriate for assessing species limits in, for example, petrels or owls?

A useful way to visualize the severity of this problem is to look at the cover of CDH14, which presents a tree that helpfully illustrates the proposed phylogenetic relationships of the 100+ families represented in the tree. However, the case for elevating 462 subspecies to species rank (a 12% increase in non-passerine species) is based on a scheme that incudes data from only three of those 100 families, two of which are on the same minor branch of the tree. Even the remaining 55 passerine species pairs from the passerine families not shown on the cover are themselves highly skewed phylogenetically, with 11 (20%) Amazonian antbirds (Thamnophilidae) and another 20 (36%) Nearctic or Palearctic migratory oscines. Consequently, the input to the T10 scheme used to justify elevation to species rank in complex groups such as albatrosses, hummingbirds, owls, and parrots received not a single data point from the orders or even major radiations to which those birds belong, but a remarkable 20% of that same input comes from Amazonian antbirds. Even application of the passerine sample to just the Passeriformes would require strong evidence that these biases do not affect the outcomes.

Collar et al. (2016) portray my published praise for the Isler et al. (1998) comparative framework for assigning species rank in the Thamnophilidae as damning contrast to my criticism of their superficially similar framework. This further exposes that Collar et al. (2016) do not recognize the importance of phylogenetic context. In contrast to CDH14, Isler et al. (1998) applied their system only to a single family, Thamnophilidae. In further contrast to CDH14, Isler et al. (1998) treat evidence for gene flow at contact zones as support for subspecies rank, as in any sensible application of the Biological Species Concept, whereas T10 and CDH14 somehow interpret rampant gene flow as evidence for separate species rank, which leads to the next major problem.

PARAPATRY

Parapatry without evidence for free gene flow is an automatic indicator of species rank in any classification. What better evidence could one get than no gene flow at the contact zone? In response to my comments on the importance of parapatry, rather than admit parapatry is *not* just another character (given a point value of 3 in their scheme, i.e., leaving four additional points needed for species rank), but instead sufficient evidence on its own for species rank, Collar et al. (2016) retreat to stating that actual determination of true parapatry is difficult and refer readers to the “entire section” (252 words) on the topic in an Appendix of T10. However, they do not acknowledge that (1) despite these obvious difficulties, they have scored an unknown number of pairs as parapatric (otherwise it would not have been included in the scoring system), and (2) they still treat this only as additional points toward treating the taxa as separate species rather than treating it as sufficient evidence for species rank, thus totally missing the point of my criticism. Further, their sudden plea for rigor in determining parapatry contrasts dramatically with the arbitrary, assumption-laden thresholds set for virtually every other parameter in the T10 scheme. Collar et al. (2016) present a *Hippolais* contact zone as their worst-case scenario of the difficulties in assessing true parapatry for two taxa that differ only slightly in plumage, with the following statement: “How would the relationship even be noticed in poorly known regions and for adjacent taxa that are still more lookalike and soundalike? In addition, what would be the chances that they never interbreed?” Yet the two *Hippolais* have been recognized as distinct species since 1817, and the width of that zone is roughly 50–100 km (Secondi et al. 2003; Fig. 1). Further, the point concerning “never interbreed” is hopefully some sort of lapsus on the part of Collar et al. (2016) because no modern application of the BSC would change taxon rank on the basis of discovery of occasional hybrids, which are irrelevant to species rank. What counts is whether almost all individuals from the contact zone appear to have intermediate characters, indicating a breakdown of barriers to gene flow, rather than the presence of occasional hybrids. Even in poorly sampled regions, there may be sufficient sampling at the scale of, say, 50 km to hint at the phenotypic distribution of characters in the contact zone. So, at least for cases where there are no obvious biogeographic discontinuities between taxa and geographic sampling is also incomplete, why not use some value centered around the lower limit of widths of narrow hybrid zones for treating taxa tentatively as parapatric species *if sampled as close as that lower limit without any sign of hybridization*? This would represent a testable hypothesis, open to revision with additional data.

MISSING THE POINT

Collar et al. (2016) attempt to counter my point concerning the resemblance of their system to long-discredited numerical taxonomy by pointing out my strong support for the Isler et al. (1998) system, which *in part* uses a numerical scale to assign species rank. This attempt again succeeds only in revealing that Collar et al. (2016) do not truly understand the differences between their scheme and that proposed by Isler et al. (1998). In addition to the difference in controlling for phylogeny pointed out above, Isler et al.'s (1998) system focuses entirely on the vocal differences known from playback trials with the thamnophilids to be associated directly with species recognition (which is why 20 of the 58 species pairs in the T10 system are from the Thamnophilidae); it is *only then* that any scoring system comes into play. In contrast, T10 assumed that all of the characters they incorporate into their scoring system *might* have an effect on species discrimination without knowing in any particular case which ones might be in operation and with weighting schemes favoring characters (as in numerical taxonomy) known to have greater importance in birds in general (although not necessarily in each family to which the scheme is applied). Thus, in the T10 system, positive points toward the threshold of seven points can be accumulated based on characters such as "a slight different wash or suffusion to all or any part of any area" or "an innate habit," whereas in Isler et al. (1998), these are irrelevant. Thus, the two systems clearly differ.

Remsen (2015) criticized CDH14 for using 200 km as the arbitrary cutoff point between "broad" and "narrow" hybrid zones as having no biological justification. Collar et al.'s (2016) response was to point out that the source for the 200-km break was explained in T10 as the average of 23 hybrid zones tabulated by Price (2008) and they concluded that the 200-km cutoff "is emphatically not without 'justification' [or] biological rationale." Again, they have missed the point. Pointing out the source of their calculations is obviously not the same as providing biological rationale for the arbitrary break. Price (2008) did not characterize hybrid zones as "broad" or "narrow" on this basis, nor do Collar et al. (2016) even attempt to explain why an arbitrary cutoff on a continuum of variation is biologically justifiable. Dodged completely is the important point that their "broad" vs. "narrow" cutoff should at least be adjusted for overall range sizes. In other words, consider the different implications of a hybrid zone 20-km wide between two taxa with range widths of 200 km vs. 2000 km. Keep in mind that the difference between "broad" and "narrow" is a difference of one point in the Tobias et al. (2010) scheme in which only six more points are needed to reach species rank.

In my original review, I pointed out that CDH14 adopted a split of two species of *Larus* gulls based on genetic evidence despite the fact that (1) the split would not have been adopted using their own point system, and (2) their own statement that "acceptance of the splits or lumps based solely on mtDNA cannot be regarded as robust." The defense of this treatment by Collar et al. (2016) is to cite the statement in CDH14 that this case represents "one of the most complex challenges in systematic ornithology . . .," as if that statement is sufficient reason to go ahead and endorse a controversial taxonomy, even provisionally, rather than wait for definitive data that their own assessment would seem to demand. Collar et al. (2016) then refer again to the same four studies of *Larus* mtDNA variation to support their treatment without acknowledging that those papers produced mtDNA gene trees, yet they treated them as species trees. Collar et al.

(2016) certainly must know that large species in the genus *Larus* hybridize frequently and are thus the epitome of candidates for gene tree/species tree conflicts. Thus, Collar et al. (2016) have missed the point, namely that the CDH14 decision contradicts their explicit position of adopting changes based solely on mtDNA gene trees. They also report that a volume that I co-authored (Dickinson and Remsen 2013) adopted this split, but Collar et al. (2016) could not know that this was one of hundreds of executive decisions made in that volume with which I disagreed. For my personal evaluation of this specific case, see my publicly available comments (fourth from bottom) at http://www.gizard.org/nacc/proposals/comments/2007_B_comments_web.html. The most important point illustrated in this example is the need for adding to the BLI/HBW team someone with expertise in interpreting gene-based phylogenetic data.

REHABILITATING A SYSTEM FOR RAPID ASSESSMENT OF TAXONOMIC RANK

I recommend the following changes to the T10 and CDH14 scheme for it to have credibility with the scientific community:

(1) Restrict its application to cases of allopatric taxa for which there are no other data available and species rank was changed by the Peters Check-list series from the prior status quo without even a phrase of published rationale.

(2) Make some attempt to control for phylogenetic effects by creating a comparative database for each family based on "knowns" in terms of interaction at secondary contact zones, and apply the results from those knowns (in terms of free gene flow or lack of it) to the unknowns (the sample of unstudied allotaxa in Recommendation 1).

(3) Create a separate system for scoring parapatric taxa.

   ○ If the taxa are truly parapatric, then there is no need to go further; the phenotypic signal from the contact zone should provide a reliable first approximation of the degree of gene flow between the taxa. If available data indicate that almost all individuals at the contact zone, regardless of width, are intermediate to some degree in phenotype, then a safe assumption is that the two gene pools are freely communicating. The chance of finding any pure parental types after only 10 generations of random mating and no selection against hybrids is roughly only 0.001% (without immigration). Obviously, an intensive gene-based study would be ideal, but under the forgiving umbrella of rapid assessment, using phenotypic characters and their individual and geographic distribution provides a testable hypothesis.

   ○ If parapatry is likely, but nonetheless uncertain, then create some arbitrary rules for treating them as truly parapatric (or not) based on distances between nearest samples. Here is where width of known hybrid zones can be used to make objective standards. Again, this is in the spirit of the urgent rapid assessment and would be subject to revision, obviously, with additional data.

(4) In revising the scheme, involve scientists whose primary research area is avian speciation. The current BirdLife team consists of evolutionary ecologists and conservation biologists with formidable knowledge of bird biology and research experience in vocal aspects of the speciation process. However, the flaws in their scheme reveal the absence of input from those whose primary research expertise is in the area of contact zones, geographic variation, and speciation, particularly from the standpoint of molecular genetics.

An overhaul of the current system is badly needed for a globally important conservation organization such as BirdLife International to retain the support of the thousands of professional and knowledgeable amateur ornithologists who support in spirit the mission of the organization. Doggedly adhering to the currently flawed T10 system will only distance BLI and HBW from professional ornithology. Collar et al. (2016) proudly point to preliminary positive reviews of the T10 system, yet conspicuously omit mentioning that no one but BLI/HBW has adopted it. It is time for Collar et al. (2016) to acknowledge that its fundamental flaws might be the reason.

LITERATURE CITED

COLLAR, N. J., L. D. C. FISHPOOL, J. DEL HOYO, J. D. PILGRIM, N. SEDDON, C. N. SPOTTISWOODE, AND J. A. TOBIAS. 2016. Toward a scoring system for species delimitation: a response to Remsen. Journal of Field Ornithology 87: 104–110.

DICKINSON, E. C., AND J. V. REMSEN, Jr. (eds.). 2013. The Howard & Moore complete checklist of the birds of the world, vol. 1, 4th ed. Aves Press, Eastbourne, UK.

GRUDZIEN, T. A., W. S. MOORE, J. R. COOK, AND D. TAGLE. 1987. Genic population structure and gene flow in the Northern Flicker (*Colaptes auratus*) hybrid zone. Auk 104: 654–664.

ISLER, M. L., P. R. ISLER, AND B. M. WHITNEY. 1998. Use of vocalizations to establish species limits in antbirds (Passeriformes: Thamnophilidae). Auk 115: 577–590.

MOORE, W. S. 1987. Random mating in the Northern Flicker hybrid zone: implications for the evolution of bright and contrasting plumage patterns in birds. Evolution 41: 539–546.

———, AND J. T. PRICE. 1993. The nature of selection in the Northern Flicker hybrid zone and its implications for speciation theory. In: Hybrid zones and the evolutionary process (R. G. Harrison, ed.), pp. 196–225. Oxford University Press, Oxford, UK.

PRICE, T. D. 2008. Speciation in birds. Roberts & Company, Greenwood Village, CO.

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

———. 2015. [Review of] HBW and BirdLife International Illustrated Checklist of the Birds of the World. Non-passerines (N. J. Collar and J. del Hoyo, eds.), vol. 1, 903 pp. Journal of Field Ornithology 86: 182–187.

SECONDI, J., V. BRETAGNOLLE, C. COMPAGNON, AND B. FAIVRE. 2003. Species-specific song convergence in a moving hybrid zone between two passerines. Biological Journal of the Linnean Society 80: 507–517.

TOBIAS, J. A., N. SEDDON, C. N. SPOTTISWOODE, J. D. PILGRIM, L. D. C. FISHPOOL, AND N. J. COLLAR. 2010. Quantitative criteria for species delimitation. Ibis 152: 724–746.