Performance of a points-based scoring system for assessing species limits in birds

ABSTRACT

Species are fundamental to biology, conservation and environmental legislation, yet there is often disagreement on how and where species limits should be drawn. Even sophisticated molecular methods have limitations, particularly in the context of geographically isolated lineages or inadequate sampling of loci. With extinction rates rising, methods are needed to assess species limits rapidly but robustly. Tobias *et al.* (2010) devised a points-based system to compare phenotypic divergence between taxa against the level of divergence in sympatric species, establishing a threshold to guide taxonomic assessments at a global scale. The method has received a mixed reception. To evaluate its performance, we identified 397 novel taxonomic splits from 328 parent taxa made by application of the criteria (in 2014–2016) and searched for subsequent publications investigating the same taxa with molecular and/or phenotypic data. Only 71 (18%) of these novel splits have since been evaluated by independent studies, suggesting that the criteria accelerated taxonomic decisions for this sample by up to 33 years overall. In the evaluated cases, independent analyses explicitly or implicitly support species status in 62 (87.3%) of 71 splits, with the level of support increasing to 97.2% when excluding subsequent studies limited only to molecular data, and reaching 100% when the points-based criteria were applied using recommended sample sizes. Despite the training set used to calibrate the criteria being heavily weighted toward passerines, splits of passerines and non-passerines received equally strong support from independent research. We conclude that the method provides a useful tool for quantifying phenotypic divergence and fast-tracking robust taxonomic decisions at a global scale.

INTRODUCTION

Species taxonomy underpins much of biological research, with the establishment of stable and globally standardised species limits being particularly critical for macroecology, macroevolution and the setting of conservation priorities. Alas, taxonomic stability and standardisation have proved elusive for all major taxa, especially birds. The waning and waxing of the number of bird species recognised by global authorities over the past century have reflected, first, a major fluctuation in the accepted definition of species and, second, a rapid expansion in knowledge and data relating to the biological characteristics of avian taxa around the world. From a high of 18,939 (Sharpe 1899–1909), the number of bird species recognised in published world checklists fell in under forty years to a mere 8,616 (Mayr 1946), but now totals between 10,175 (Christidis *et al.* 2018) and 11,158 (HBW/BirdLife International 2020) species.

The upward trend in the number of recognised bird species in recent decades does not reflect the discovery of new species so much as the redrawing of species boundaries based on new information and new ways of processing old information. Rates of taxonomic change have been driven in part by legions of birdwatchers and sound-recordists collecting data on distributions and vocal signals, as well as by museum researchers examining larger samples of preserved material. An even more significant role has been played by molecular biologists and phylogeneticists uncovering the evolutionary relationships among taxa and developing new tools to examine species boundaries. Nevertheless, the interpretation and use of the information, new and old, has been far from consistent. Different world and regional lists have used different methods to gauge the validity of the myriad taxonomic judgements that derive from the continuing cascade of new information, resulting in a divergence of listings that has been characterized as 'taxonomic anarchy' (Garnett and Christidis 2017).

To some extent this is a matter of preferred 'species concept', the multiple forms of which are themselves the most salient evidence of 'anarchy' in taxonomy (for ornithology see Haffer 1992, 1997). Species concepts come in various guises, but in the context of avian taxonomy the choice is often between the 'phylogenetic species concept' (PSC), which espouses monophyly as its key criterion, and the 'biological species concept' (BSC), which makes reproductive incompatibility its central test (e.g. Winker *et al.* 2007). Although both approaches have advantages and limitations, a wholesale switch to the PSC is currently viewed as problematic from the perspective of standardising species lists (Collar 1997, 2018, Johnson *et al.* 1999), not least because PSC-based assessments may double or treble the

65   number of recognized bird species (Barrowclough *et al.* 2016). Therefore, world and regional
66   lists still adhere, at least nominally, to the BSC, building upward and outward from the base
67   provided by Peters and successors (1931–1986). Even so, the subjectivity involved in
68   deciding the rank of allopatric taxa under the BSC remains a serious challenge, leading to
69   ever-increasing disparities between the various world lists of bird species (Garnett and
70   Christidis 2017).
71   The problem of 'subjectivity' of the BSC in assigning rank to allopatric taxa was
72   treated as unfounded ('allopatric populations can be assessed objectively') in a robust
73   rejection of the PSC by Johnson *et al.* (1999), whose argument however rested entirely on
74   emerging technological capacities to analyse vocalisations, displays and genetic sequences.
75   Their paper offered no way forward in matters of morphological (plumage and size)
76   differences, and failed to reflect on the complexities of acoustic analysis in the context of
77   learned or innate vocalizations, or of genetic analysis where outcomes are obscured by
78   incomplete sampling and rapidly changing methodologies. Its concluding formulation of a
79   'comprehensive biologic [sic] species concept' could commonly be applied to taxa ranked as
80   subspecies.
81   Shortly afterwards, Helbig *et al.* (2002) proposed that species rank be assigned to
82   allopatric taxa that are 'fully diagnosable in each of several discrete or continuously varying
83   characters' and where 'the sum of the character differences corresponds to or exceeds the
84   level of divergence seen in related species that coexist in sympatry'. They further proposed
85   that 'allospecies' rank—without clarifying what this represents—be assigned to taxa that are
86   fully diagnosable by at least one character and in which the level of divergence is equivalent
87   to that found in related sympatric species. The difficulty with this formulation lies in the non-
88   specificity of 'several', the potential triviality of the diagnostic characters identified, the
89   predictable difficulty in many cases of finding sufficiently closely related sympatric species
90   with which to make comparisons, and problems in gauging levels of divergence when
91   represented by different kinds of character. Thus, they could be used to assign species rank to
92   virtually any diagnosable subspecies. Perhaps for this reason these criteria have been little
93   used, even by their own authors, who in multiple subsequent decisions on allopatric taxa in
94   their 'taxonomic recommendations for British birds' (*Ibis* volumes 144–158) did not apply
95   the comparison with levels of difference in 'related sympatric species'.
96   Recognising the weaknesses in these formulations but respecting the intention behind
97   them, Tobias *et al.* (2010) proposed what were intended to be more robust and more explicit
98   criteria by allowing for *strength* of character as well as number. These were based on an

99 exercise in which 58 pairs of closely related and morphologically similar sympatric or

100 parapatric bird species from all continents and latitudes (albeit largely focused on passerines)

101 were scored for the strength and number of their phenotypic and ecological differences. The

102 level of divergence in these universally accepted species was established as a preliminary

103 benchmark (or 'yardstick') to assess species status, a concept with a long history in

104 ornithology (Mayr 1969, Isler et al. 1998, Helbig *et al.* 2002) and championed more widely

105 as a solution to runaway taxonomic inflation under the PSC (Isaac et al. 2004, Meiri and

106 Mace 2007). A simple method was developed for converting quantitative measurements into

107 scores, and a variety of limits were placed on scoring to increase objectivity and avoid double

108 counting (Figure 1). This study suggested that a total score of 7 can serve as a general

109 threshold for the recognition of species rank in birds.

110   In a recent world checklist (del Hoyo and Collar 2014–2016; henceforth 'the

111 Checklist'), these methods (henceforth 'the criteria') were applied in cases where the

112 elevation of distinctive-looking (and/or distinctive-sounding) subspecies to species appeared

113 a possible outcome. (The Checklist also incorporated numerous revisions following other

114 sources, including many molecular studies, where the criteria were often used as a means of

115 validation: Burfield *et al.* 2017.) This list, which is updated annually using the same approach

116 (see http://datazone.birdlife.org/species/taxonomy), now underpins the taxonomy for birds on

117 the IUCN Red List, and has widespread policy impact (for example, being adopted by various

118 multilateral environmental agreements: Burfield *et al.* 2017). The criteria have experienced a

119 relatively negative reception in some quarters, yet in the decade after they appeared in print

120 no peer-reviewed paper has to our knowledge subjected them to direct analytical scrutiny.

121   We therefore assess quantitatively how well the criteria have performed in matching

122 and predicting taxonomic proposals, by comparing the results of the application of the criteria

123 during the preparation of the 2014–2016 Checklist with independent research that has been

124 published subsequently, based on analyses of both molecular and phenotypic data. We also

125 take the opportunity to address criticisms and common misunderstandings of the criteria.

126

127 **METHODS**

128

129 We identified all taxa that were recognized as species in the Checklist (del Hoyo and Collar

130 2014–2016) that were split on the basis of the criteria and not previously given species rank

131 in either the 16 volumes of 'HBW' (del Hoyo *et al.* 1992–2011) or the 2013 version of the

132 BirdLife International taxonomic checklist (BirdLife International 2013). We then excluded

133   all newly described species and any species already recognized by the IOC list

134   (https://www.worldbirdnames.org/new/) in versions prior to the publication of the

135   Checklist—with reference to v4.1 (January 2014) for non-passerines and v6.1 (January 2016)

136   for passerines—or by Howard and Moore v4.1 (2014). To avoid the potential bias of

137   sampling splits most likely to be supported, we also excluded several hundred taxa split in the

138   Checklist which had previously (since the 1950s but prior to the Checklist) been proposed by

139   other authors for treatment as species. This left 397 novel splits (excluding nominate

140   subspecies) from 328 'parent' taxa that were proposed by the Checklist based on the criteria.

141   To identify which of these had been subsequently and independently assessed by

142   others, we compiled a library of papers relating to the taxonomy of the world's birds

143   published from 2014 to October 2020, including a comprehensive survey from 2014–2019

144   provided by J. L. Copete of Lynx Edicions of some 120 journals that published over 500

145   papers and articles relevant to bird taxonomy. This survey was supplemented by J. V.

146   Remsen's regular circulation of new titles and by our own regular searches. The resulting

147   inventory was thus likely to have been close to a comprehensive list of the avian taxonomic

148   literature post-dating the Checklist up to and including 2019, with a less comprehensive

149   collection of further studies published up to October 2020. From this we then sought, through

150   a rapid review of the assembled material, to identify those taxonomic studies by other authors

151   that simultaneously or subsequently, and independently, assessed the parent taxa that had

152   been split by application of the criteria in the Checklist. Such studies included both explicit

153   assessments of species limits, either at the level of individual species or across higher

154   taxonomic levels, and studies that presented taxonomic data, for example in the form of

155   phylogenies, without making explicit recommendations regarding species limits. To

156   retrospectively assess the extent to which our methods are likely to have missed relevant

157   papers, we used a two-stage search process. First, we randomly selected 50 novel splits from

158   different parent taxa for which we had not found a subsequent independent study, and

159   searched on the scientific name of the parent species, and the term 'taxonomy', in Google

160   Scholar. We then searched the list of results using the species name of the split, and examined

161   all remaining results for evidence that the taxon or taxa had been assessed. We then calibrated

162   the results of this process by repeating it exactly for 20 randomly selected splits from

163   different parent taxa for which we had identified subsequent studies.

164   We scored the outcome of each such study according to whether it provided evidence

165   that (1) did not support the split, (2) implicitly supported the split, for example by showing a

166   phylogeny or indicating significant vocal or morphometric differences that support the split

167   but without making a direct taxonomic declaration in its favour, or (3) explicitly supported

168   the split by presenting evidence in its favour and using that evidence to propose that the same

169   taxonomic division be made. Previous studies have criticized the criteria on the basis that the

170   threshold of 7 for species status was calibrated on a sample heavily biased towards

171   passerines, and therefore may not be relevant to non-passerines (Remsen 2016). To assess

172   whether the criteria performed differently in passerines and non-passerines, we quantified

173   support for these taxonomic groupings separately.

174   Full application of the criteria involves the use of measurements from at least 10

175   individuals when calculating effect size of morphometric or vocal differences between taxa.

176   However, when applying the criteria, assessments of vocal differences or morphological

177   measurements were often based on samples of fewer than 10 individuals per taxon, owing to

178   a shortage of available material. To assess whether splits were more likely to be supported if

179   they were based on more robust sample sizes, we scored sampling (or 'robustness') as 'Low'

180   = split resulting from scores assigned on the basis of a combination of plumage, ecology,

181   geography, morphology (but with <10 specimens per taxon measured) and voice (with <10

182   individuals typically sampled); 'Medium' = split resulting from scores assigned on the basis

183   of a combination of fully sampled plumage, ecology, geography and morphology (with ≥10

184   specimens per taxon measured) but with under-sampled vocal data (<10 individuals

185   sampled); or 'High' = split resulting from scores assigned on the basis of plumage alone or

186   any combination of plumage, measurements (with ≥10 specimens per taxon measured),

187   ecology and geography (i.e. scores for voice not needed, thus no parameters were based on

188   <10 individuals per taxon). The criteria for 'High' robustness meet the samples sizes

189   recommended by Tobias *et al.* (2010).

190

191   **RESULTS**

192

193   We identified 328 parent taxa for which application of the criteria led to the split of one or

194   more previously unrecognized species in the Checklist, totalling 397 splits. Of these, we

195   found subsequent studies of 60 (18.3%), which between them accounted for 71 splits (17.9%)

196   made in the Checklist through application of the criteria (Table 1). Searches in Google

197   Scholar for a random sample of 50 of the 268 parent taxa for which we found no subsequent

198   study returned only one paper relevant to our purposes, and that was published outside our

199   search period (in December 2020). In contrast, when applied to a random sample of 20 parent

200   taxa for which we identified subsequent studies, the same search protocol successfully

201 located those studies in 17 cases, usually as the first item in the list of search results. The

202 three cases that were missed either used a different genus name (two instances), or the same

203 genus name was abbreviated to a single letter where the target taxon was mentioned in the

204 paper, so the search term was not located. Taking the probability of a paper being missed by

205 the search protocol as $1–(17/20) = 0.15$ (with exact binomial 95% CL of 0.05–0.36), we

206 estimate the probability of returning no search results from a random sample of 50 of the 268

207 parent taxa for which no studies were found if our sample of 71 were 90% complete as

208 $0.15^{((71*(100/90))-71)*(50/268)} = 0.06$ (95% CL: 0.012–0.22). Thus we can be 78–99% confident

209 that our sample contains 90% or more of relevant studies. Furthermore, we see no reason why

210 studies that we missed should be systematically different from those we found in terms of

211 their support or otherwise for splits made using the criteria.

212   Of these 71 splits, 9 (of 6 parent taxa) received no support from subsequent research

213 (although in 2 cases they were identified as separate 'conservation units'), 19 (of 16 parent

214 taxa) received implicit support from subsequent research, and 43 (of 38 parent taxa) received

215 explicit support (Table 1, Figure 2). Thus, the overall concordance rate was 87.3% (62 of 71

216 splits), with no difference between passerines and non-passerines in the extent to which splits

217 made by application of the criteria received subsequent independent support (Figure 2).

218 However, in only 13 cases were the original splits based on the sample size of 10 or more

219 individuals per taxon recommended under the criteria, owing to paucity of accessible

220 specimens or recordings. In the case of these 13 splits, all (100%) were implicitly or

221 explicitly supported by subsequent independent research (Figure 2). In addition, there was a

222 significant association between the degree of support and the methods used in subsequent

223 analyses (genetic data only, phenotypic data only, and integrated genotype/phenotype data:

224 Fisher's exact test, $P < 0.001$; Figure 2). This was because for 8 of the 9 splits that received

225 no support, subsequent evaluation had been based on analyses with genetic data only. When

226 excluding cases where subsequent studies did not use any genetic methods, the concordance

227 rate rose to 95%, and when excluding genetics-only studies it rose again to 97.2%. Finally,

228 concordance was 100% among studies that used an 'integrative taxonomy' approach,

229 combining genetic, phenotypic and other metrics (Figure 2).

230   Of the 196 taxa that were scored against the criteria and not split on the basis of the

231 resulting scores, we only found 6 that were independently assessed subsequently; in 3 cases

232 the decision not to split was supported and in 3 cases the subsequent analyses proposed splits.

233 Although sample sizes were small, the proportion of splits based on the criteria that received

234 subsequent independent support (explicit or implicit) was significantly higher than the

7

proportion of non-splits that received subsequent support (62/71 *vs* 3/6; Fisher's exact test, P

< 0.02), suggesting that the criteria may be conservative in their likelihood of producing

splits with respect to the wider taxonomic literature. Furthermore, the proportion of criteria-

based splits that went on to be independently assessed (60/328 parent taxa; 18.3%) was

significantly higher than the proportion of criteria-based non-splits that received subsequent

independent taxonomic attention (6/196; 3.1%; Fisher's exact test, P < 0.0001), suggesting

that research attention was more focused towards taxa split by application of the criteria.

During the preparation of the Checklist, we also applied the criteria to 145 taxa

proposed by other authorities to be treated as species largely or wholly on the basis of

molecular evidence. Of these, 120 (82.8%) qualified as species by application of the criteria

and were treated as such in the Checklist, while 25 (17.2%) were treated as subspecies

because, on the available evidence (acknowledging that analysis of additional data, especially

acoustic, might alter these conclusions), they scored less than 7 using the criteria. It is notable

that none of the 25 rejected cases had sample sizes classed as 'high'.

A key rationale for the development of the criteria was operational speed. Taking

2015 as the average publication date of the two Checklists, in the following six years up to

2020 inclusive, 60 of our restricted sample of 328 parent taxa that were split by application of

the scoring method were subsequently independently analysed. On this basis, if we

simplistically assume the same rate of taxonomic investigation and concordance (i.e. 60/6 =

10 parent taxa split per year), it would take until 2047 for the remaining 268 parent taxa to be

subject to independent scrutiny. Use of the criteria to verify splits in the Checklist therefore

brought forward taxonomic revisions by up to 33 years, with the true figure likely to be

substantially larger considering the increased research attention on splits in the Checklist (see

above), and the hundreds of other previously proposed splits that were assessed but not

included in our sample (see Methods).

DISCUSSION

The key finding of this review is that the criteria of Tobias *et al.* (2010) produce decisions on

biological species limits in birds that conform well with decisions subsequently reached by

other researchers using various other methods. Splits of species first proposed by application

of the criteria were ten times more likely to receive support from subsequent independent

study than not. The context where the criteria matched least well with the conclusions of

independent analyses is when those subsequent studies were based solely on molecular data.

269 However, this lower level of support can be partly explained by the higher proportion of

270 genotype-only studies adopting an entirely different species concept (PSC), sometimes

271 leading to taxonomic recommendations at odds with the BSC. Concordance with the

272 taxonomic conclusions or implications of previous studies was 100% when the criteria were

273 applied based on better sampling (see Figure 1), or when evaluated with integrative studies

274 based on a range of data sources (Figure 2). This of course does not imply that the criteria can

275 achieve total concordance with other treatments, particularly given numerous potential

276 sources of error and bias underlying the scores, as well as the inherent subjectivity of all

277 taxonomic classification systems (Tobias et al. 2010). Nonetheless, the degree of agreement

278 among studies is highly encouraging, and suggests that the criteria can deliver robust

279 taxonomic assessments in a short time-frame.

280 Our observation that performance improves with more extensive sampling of

281 individuals in morphometric and vocal analyses is likely to be caused by two main factors.

282 First, larger samples reduce error and uncertainty in effect sizes, and second, they also reduce

283 the problem of bias caused by individual variation or contextual differences (e.g. songs

284 sampled in different seasons or settings). Another sampling issue justifiably raised in a

285 previous review of the criteria (Remsen 2016) was that the species pairs used to derive the

286 scoring threshold were dominated by passerines. This bias in sampling would obviously be

287 problematical if it meant the criteria worked less well for non-passerines. However, we found

288 no evidence of this effect because splits of non-passerines were equally likely to receive

289 subsequent independent support.

290 The criteria are designed for fast-tracking taxonomic decisions and are therefore

291 relatively basic by design, as discussed by Tobias et al. (2010). We do not believe that there

292 is anything 'magic' about the score of 7, as defined in Figure 1, merely that it reflects a level

293 of differentiation between taxa that correlates with species limits assigned through other

294 methods. There is a tendency among ornithologists to assume that highly sophisticated

295 genomic analyses must be able to delimit species more accurately than any method relying

296 largely on phenotypic divergence, but this assumption is risky in the case of allopatric taxa

297 for which monophyly can simply reflect population structure (arising from reduced gene flow

298 between spatially isolated populations of the same species) (Tobias et al. 2010, 2020). The

299 criteria counter this problem by defining a minimum threshold of phenotypic divergence

300 associated with sympatric or parapatric species pairs. Our results suggest that even in their

301 current form the criteria offer a useful tool for making largely accurate taxonomic

302 classifications across numerous species in a short time-frame.

303

PROMISE AND PITFALLS OF GENETIC EVIDENCE FOR REPRODUCTIVE
ISOLATION

306

In the decade since their inception, the criteria have received plenty of criticism, often on the basis of misunderstandings about their goals and methods (Collar et al. 2016). A common misconception is that the approach taken is anti-molecular, and molecular biologists have been understandably sensitive to the fact that genetic information were not incorporated as a quantifiable component into the criteria (e.g. Collinson *et al.* 2017). However, far from ignoring genetic evidence, Tobias *et al.* (2010) explicitly prioritise genetic data when populations are in contact (Figure 1) and emphasise that the criteria are designed to accommodate genetic distances as soon as their relevance to species limits under the BSC are better understood.

The reason genetic measures were not integrated at the outset is because simplified molecular metrics – along the lines of mtDNA divergence – are difficult to interpret in the context of species limits (Winker *et al.* 2007). The reasons for this problem are widely reported, including heterogeneous rates of molecular evolution and gene flow among currently or historically parapatric taxa, which can pose serious difficulties for species delimitation (Tobias et al. 2020, Jiao and Yang 2021). When secondary contact occurs before reproductive isolation is complete, hybridization can lead to anything from zero gene flow to complete merger of gene pools (e.g. Kearns *et al.* 2018), or genome-wide introgression with species limits maintained by very few 'barrier loci', often on the sex chromosomes (e.g. Toews *et al.* 2016). In this context, introgression of the mitogenome can periodically reset mtDNA divergence to zero during the process of speciation with gene flow, even accounting for unexpectedly low mtDNA divergence among well-established species (Irwin *et al.* 2009, Rheindt and Edwards 2011, Tobias *et al.* 2020, Miller *et al.* 2021).

While it is often argued that genetic information provides greater reliability in taxonomic decisions, any system over-reliant on molecular evidence will also suffer instability. For example, Martens et al. (2008) described a new species – Alpine Leaf-warbler *Phylloscopus occisinensis* – on the basis of divergent mtDNA. However, this was later identified as a case of 'deep mitochondrial divergence' within populations of *P. affinis* potentially caused by hybridization with a now-extinct congener, which left the 'ghost of introgression past' in the genes of the colonizing population (Zhang et al. 2019). When assessed under points-based criteria, *P. occisinensis* did not qualify for species status because

337   it was undiagnosable on vocal or plumage characters. This outcome was uniquely stable

338   because all other major taxonomic checklists adopted the split of *P. occisinensis*, apparently

339   in error.

340   In other cases, instability caused by deference to genetic evidence stems from the

341   ever-changing landscape of molecular methods and evolutionary modelling. As methods and

342   datasets improve, reversals of earlier judgements can diminish confidence in work published

343   only a few years before. Contradictory assessments of the genetic differences between

344   Common *Apus apus* and Pallid Swifts *A. pallidus* (Päckert *et al.* 2012, Pellegrino *et al.* 2017)

345   and of the validity of *Heliangelus zusii* (Kirchman *et al.* 2010, Pérez-Emán *et al.* 2018) are

346   cases in point. Overall, there is an emerging view that the use of mtDNA alone, an approach

347   used in most of the earlier genetic studies of birds, is insufficient to determine taxonomic

348   relationships reliably: Drovetski *et al.* (2018) 'caution against the out-of-hand dismissal of

349   traditional taxonomy in cases when mtDNA appears to contradict it, regardless of how strong

350   the support of geographically coherent clades in the mtDNA gene tree might be'.

351   None of this is intended to downplay the crucial contribution of molecular evidence to

352   systematic revision and species delimitation in birds. Both genotypic and phenotypic

353   evidence have clear strengths and limitations, and they clearly need to be considered in

354   conjunction as complementary components of 'integrative taxonomy' (Winker et al. 2009,

355   Padial *et al.* 2010). Many impressive studies have taken up the challenge of bringing multiple

356   lines of evidence to bear on taxonomic judgements (e.g. Alström et al. 2008, 2018a, Cadena

357   and Cuervo 2010). These are the gold standard in terms of integrative taxonomy, although

358   some cases appear to give heavier weighting to genotype. Hosner *et al.* (2018), for example,

359   invoked 'operational criteria' for deciding species rank that involve the congruence of '(1)

360   well-supported monophyly of geographic clades, (2) significant genetic differentiation, as

361   identified by a coalescent model, and (3) fixed plumage and morphological differences'. In

362   this formulation, taxonomy is integrative but not exactly balanced: as long as the first two

363   genetic criteria are met, morphological differences, however tiny, can trigger species rank.

364   Even without such differences the first two criteria identify 'cryptic lineages' which *may* be

365   species, although the authors admit 'their genetic distinctiveness could be an artifact of strong

366   population structure'. If so, however, why should this possibility not equally extend to splits

367   that are only marginally distinct in morphology?

368   In a related study of Bornean birds, Moyle *et al.* (2017) found significant genetic

369   differences between lowland and upland representatives of three taxa with apparent

370   'elevational parapatry', arguing that these consequently merited recognition at species level,

371   and concluding with the remark that 'any species concept that attempts to predict

372   interbreeding potential simply on the basis of perceived morphological differences (Tobias *et*

373   *al.*, 2010) likely underestimates species diversity'. While this may be the case, we think the

374   criteria would only overlook a very small proportion of parapatric cryptic species acceptable

375   under the BSC simply because the score for parapatry is high (3), and when added to

376   consistent vocal differences, often triggers species status (Figure 1). Given how regularly this

377   simple fact is overlooked, we suspect that some molecular biologists who have dismissed the

378   criteria may have never attempted to apply them fully in practice.

379   A classic example of over-hasty dismissal is provided by a recent study establishing

380   *Catharus maculatus* as a species distinct from *C. dryas* (Halley *et al.* 2017). The authors

381   argued that their split would fail under the 'yardstick' criteria used in previous studies, yet

382   also reported that the two taxa 'are 100% diagnosable in genetic, vocal, morphometric, and

383   plumage characters'. These findings suggest that *C. maculatus* and *C. dryas* would be treated

384   as species under all three permutations of yardstick criteria developed for birds (Isler et al.

385   1998, Helbig *et al.* 2002, Tobias *et al.* 2010). Indeed, when the data from Halley *et al.* (2017)

386   are applied to the Tobias et al. (2010) criteria, *C. maculatus* emerges with a score of >10,

387   well above the threshold set for species rank. In some of these cases, we suspect that

388   misinterpretation may arise from an assumption that any taxon not split in the Checklist failed

389   to meet the 7-point threshold, whereas in many cases a lack of splitting simply means that the

390   case was not investigated or the relevant data were not available. The case of *Catharus*

391   *maculatus*, for example, was not scored for the Checklist and so does not appear in Table 1 of

392   this paper.

393

394   **LIMITATIONS AND FUTURE DIRECTIONS**

395

396   Species delimitation is to some extent inherently arbitrary, and disagreement about the

397   conclusions of any taxonomic system is therefore unavoidable (Hey et al. 2003, Winker et al.

398   2007). With regard to taxonomic changes in the Checklist, a recurring point of contention

399   involves the assignment of species status to taxa joined by hybrid zones (e.g. Donegan et al.

400   2015). The criteria are designed to classify such cases as species even when lineages

401   hybridise freely across a broad zone, as long as lineages appear to retain highly divergent and

402   stable phenotypes on either side of the zone, as explained with reference to examples in

403   Collar et al. (2016). We consider our approach to be supported by evidence that species limits

404   can be maintained in genetically near-identical species by relatively few barrier loci (i.e.

405   islands of genomic differentiation), such as those coding for diagnostic plumage differences

406   in *Vermivora* warblers (Toews *et al.* 2016). In other cases, new information about the level of

407   phenotypic divergence, or the extent of genetic introgression, or indeed the width of the

408   hybrid zone in relation to the overall range of taxa, may support remerging of taxa split by the

409   criteria (Céspedes-Arias et al. 2021). However, our rationale for conferring species status to

410   hybridizing taxa with high scores of phenotypic divergence remains unchanged.

411   Another widely repeated criticism is that the criteria depend on 'subjective'

412   assessments of character differences (e.g. Martens and Bahr 2016). For instance, Hosner *et al.*

413   (2018) claimed that the criteria are 'subject to individual interpretation, and often result in

414   conflicting limits drawn from differing data sources', although they provided no evidence to

415   support the latter point and, as they noted, their own proposed criteria require 'individual

416   interpretation' to decide the taxonomic status of cryptic lineages. To some extent, this echoes

417   earlier species concept debates, with cladists criticising the BSC for relying on subjective

418   assessments, despite equivalent levels of subjectivity inherent in deciding species limits under

419   the PSC (Johnson *et al.* 1999, Winker *et al.* 2007, Tobias et al. 2010). Nonetheless, we

420   acknowledge that some degree of subjectivity is impossible to eliminate from the criteria, and

421   that this issue may be accentuated because most of the scores in the first round of taxonomic

422   evaluations for the Checklist were made by a single observer (NJC). Further steps should be

423   taken to refine the process in this regard, including averaging across scores from multiple

424   observers whenever possible. Another solution to the problem of subjectivity is transparency.

425   Not only is the scoring of character differences under the criteria regulated by explicit

426   guidelines, but the scores for individual characters are reported along with any underlying

427   quantitative data, including samples of vocal and morphological measurements used to

428   generate effect sizes. To ensure that the scores can be checked and challenged where

429   necessary, these datasets are publicly available for examination (see Data depository).

430   Some published suggestions relating to the criteria are constructive. The use and

431   treatment of effect sizes has been debated (Donegan 2018), including the suggestion that the

432   approach be abandoned entirely because of problems associated with delimiting species on

433   the basis of central tendency in phenotypic data (Cadena et al. 2018). In addition, it has been

434   suggested that the procedure used to assign and add scores violates elements of measurement

435   theory (for discussion of which, see Houle et al. 2011). We agree that these issues warrant

436   attention, and that alternative models and procedures should be examined in cases where

437   larger phenotypic datasets are available. However, we also caution against over-complicating

438   an approach designed for rapid application which already seems to work remarkably well in

439   its current format. Thankfully, not all systems are as taxonomically intractable as the

440   *Geospiza* finches investigated by Cadena et al. (2018).

441   The scoring system and associated threshold value should be applied to a wider range

442   of accepted sympatric or parapatric species, particularly non-passerines, to refine and

443   recalibrate the system. Furthermore, additional work might reveal that a wider range of

444   phenotypic characters could usefully be included in the system; spectral reflectance of

445   feathers, tarsal scutellation, osteology, egg colour, plumage of downy young and even smell

446   have all been used or proposed as ways to assess taxonomic status. Finally, even in its current

447   form, the method can provide a useful framework for quantifying phenotypic divergence in

448   studies testing evolutionary hypotheses (e.g. Campbell et al. 2016).

449

450   **CONCLUSIONS**

451   Ideally, taxonomic decisions should be based entirely on painstaking research combining

452   multiple lines of genetic and phenotypic evidence. However, with many threats intensifying

453   and bird populations rapidly declining worldwide, there is a strong argument for seeking a

454   more rapid approach – at least as a preliminary assessment – to fast-track taxonomic

455   decisions before we lose many cryptic bird species forever (Lees and Pimm 2015, Remsen

456   2016). Our findings suggest that the criteria, as applied in the Checklists, provide a

457   reasonably reliable method for achieving this goal. They have come under fire from some

458   quarters for being too 'quick and dirty', but our results suggest that they produce taxonomic

459   decisions with over 90% accuracy in a fraction of the time, and also improve the knowledge

460   base by increasing the focus of subsequent research on likely splits. The criteria therefore

461   offer a useful tool for proactively investigating hitherto neglected cases and reactively

462   evaluating taxonomic changes proposed by other analytical methods.

463

473　**Data depository:** Data used in all calculations and analyses in this paper are presented in

474　Table 1. All measurements, scores and the rationale for scores underlying all taxonomic

475　decisions made by BirdLife International in the Checklists are freely available online

476　(http://datazone.birdlife.org/species/taxonomy; acoustic

477　analyses: http://datazone.birdlife.org/userfiles/file/Species/Taxonomy/2._Acoustic_data.zip;

478　morphometric measurements:

479　http://datazone.birdlife.org/userfiles/file/Species/Taxonomy/3._Morphometric_data.zip).

480

481　**LITERATURE CITED**

482　Alström, P., P. C. Rasmussen, U. Olsson, and P. Sundberg (2008). Species delimitation based

483　　on multiple criteria: the Spotted Bush Warbler *Bradypterus thoracicus* complex (Aves:

484　　Megaluridae). Biological Journal of the Linnean Society 154:291–307.

485　Alström, P., A. Cibois, M. Irestedt, D. Zuccon, M. Gelang, J. Fjeldså, M. J. Andersen, R. G.

486　　Moyle, E. Pasquet, and U. Olsson (2018a). Comprehensive molecular phylogeny of the

487　　grassbirds and allies (Locustellidae) reveals extensive non-monophyly of traditional

488　　genera, and a proposal for a new classification. Molecular Phylogenetics and Evolution

489　　127:367–375.

490　Alström, P., P. C. Rasmussen, C. Xia, M. Gelang, Y. Liu, G. Chen, H. Zhao, Y. Hao, C.

491　　Zhao, J. Zhao, C. Yao, J. A. Eaton, R. Hutchinson, F. Lei, and U. Olsson (2018b).

492　　Taxonomy of the White browed Shortwing (*Brachypteryx montana*) complex on mainland

493　　Asia and Taiwan: an integrative approach supports recognition of three instead of one

494　　species. Avian Research 9:34.

495　Andersen, M. J., J. M. McCullough, W. M. Mauck, B. Tilston Smith, and R. G. Moyle

496　　(2018). A phylogeny of kingfishers reveals an Indomalayan origin and elevated rates of

497　　diversification on oceanic islands. Journal of Biogeography 45:269–281.

498　Barrowclough, G. F., J. Cracraft, J. Klicka, J., and R. M. Zink (2016). How many kinds of

499　　bird are there and why does it matter? PLoS ONE 11:e0166307.

500　Barth, J. M. I., M. Matschiner, and B. C. Robertson (2013). Phylogenetic position and

501　　subspecies divergence of the endangered New Zealand Dotterel (*Charadrius obscurus*).

502　　PLoS ONE 8(10): e78068. doi:10.1371/journal.pone.0078068.

503　Baveja, P., K. M. Garg, B. Chattopadhyay, K. R. Sadanandan, D. M. Prawiradilaga, P. Yuda,

504　　J. G. H. Lee, and F. E. Rheindt (2020). Using historical genome-wide DNA to unravel the

505　　confused taxonomy in a songbird lineage that is extinct in the wild. Evolutionary

506　　Applications DOI: 10.1111/eva.13149.

507 BirdLife International (2013). The BirdLife Checklist of the Birds of the World, with

508 Conservation Status and Taxonomic Sources. Version 6. Downloaded from

509 http://www.birdlife.org/datazone/userfiles/file/Species/Taxonomy/BirdLife_Checklist_Ver

510 sion_6.zip

511 Bocalini, F., and L. F. Silveira (2015). Morphological variability and taxonomy of the Blue-

512 winged Parrotlet *Forpus xanthopterygius* (Psittacidae). Revista Brasileira de Ornitologia

513 23:64–75.

514 Bock, W. J., and J. Farrand (1980). The number of species and genera of recent birds.

515 American Musuem Novitates 2703.

516 Bruxaux, J., M. Gabrielli, H. Ashari, R. Prys-Jones, L. Joseph, B. Milá, G. Besnard, and C.

517 Thébaud (2018). Recovering the evolutionary history of crowned pigeons (Columbidae:

518 *Goura*): implications for the biogeography and conservation of New Guinean lowland

519 birds. Molecular Phylogenetics and Evolution 120:248–258.

520 Buainain, N., C. P. de Assis, and M. A. Raposo (2017). Geographic variation and taxonomy

521 of the *Arremon taciturnus* (Hermann, 1783) species complex (Aves: Passerellidae).

522 Journal of Ornithology 158:631–650.

523 Burfield, I. R., S. H. M. Butchart, and N. J. Collar (2017). BirdLife, conservation and

524 taxonomy. Bird Conservation Internatational 27:1–5.

525 Cadena, C. D., and A. M. Cuervo (2010). Molecules, ecology, morphology and songs in

526 concert: how many species is *Arremon torquatus* (Aves: Emberizidae)? Biol. J. Linn. Soc.

527 99:152–176.

528 Cadena, C. D., F. Zapata, and I. Jiménez (2018). Issues and perspectives in species

529 delimitation using phenotypic data: Atlantean evolution in Darwin's Finches. Systematic

530 Biology 67:181–194.

531 Campbell, K. K., T. M. Braile, and K. Winker (2016). Integration of genetic and phenotypic

532 data in 48 lineages of Philippine birds shows heterogeneous divergence processes and

533 numerous cryptic species. PLoS ONE 11: e0159325.

534 Carmi, O., C. C. Witt, A. Jaramillo, and J. P. Dumbacher (2016). Phylogeography of the

535 Vermilion Flycatcher species complex: multiple speciation events, shifts in migratory

536 behavior, and an apparent extinction of a Galápagos-endemic bird species. Molecular

537 Phylogenetics and Evolution 102:152–173.

538 Cavarzere, V., L. F. Silveira, M. F. de Vasconcelos, R. Grantsau, and F. C. Straube (2014).

539 Taxonomy and biogeography of *Stephanoxis* Simon, 1897 (Aves: Trochilidae). Papéis

540 Avulsos de Zoologia 54:69–79.

541   Céspedes-Arias, L. N., A. M. Cuervo, E. Bonaccorso, M. Castro-Farias, A. Mendoza-

542   Santacruz, J. L. Pérez-Emán, C. C. Witt, and C. D. Cadena (2021). Extensive

543   hybridization between two Andean warbler species with shallow divergence in mtDNA.

544   *Ornithology*, ukaa065 doi.org/10.1093/ornithology/ukaa065.

545   Christidis, L., and unspecified co-workers (2018). The Howard and Moore Complete

546   Checklist of the Birds of the World, v 4.1. https://www.howardandmoore.org.

547   Clements, J. F. (2007). The Clements Checklist of the Birds of the World. Sixth edition.

548   Christopher Helm, London.

549   Collar, N. J. (1997). Taxonomy and conservation: chicken and egg. Bulletin British

550   Ornithologists' Club 117:122–136.

551   Collar, N. J. (2018). Taxonomy as tyranny. The Ibis 160:481–484.

552   Collar, N. J., L. D. C. Fishpool, J. del Hoyo, J. D. Pilgrim, N. Seddon, C. N. Spottiswoode, J.

553   A. Tobias (2016). Toward a scoring system for species delimitation: a response to

554   Remsen. Journal of Field Ornithology 87:104–115.

555   Collinson, J. M., P. Dufour, A. A. Hamza, Y. Lawrie, M. Elliott, C. Barlow, and Crochet, P.-

556   A. (2017). When morphology is not reflected by molecular phylogeny: the case of three

557   'orange-billed terns' *Thalasseus maximus*, *Thalasseus bergii* and *Thalasseus bengalensis*

558   (Charadriiformes: Laridae). Biological Journal of the Linnean Society 20:1–7.

559   Cros, E., and Rheindt, F. E. (2017). Massive bioacoustic analysis suggests introgression

560   across Pleistocene land bridges in *Mixornis* tit-babblers. Journal of Ornithology 158:407–

561   419.

562   Dai, C., F. Dong, and X. Yang (2019). Morphotypes or distinct species? A multilocus

563   assessment of two East Asian scimitar babblers (Aves, Timaliidae). Zoologica Scripta

564   49:265–279.

565   Dejtaradol, A., S. C. Renner, S. Karapan, P. J. J. Bates, R. G. Moyle, and M. Päckert (2016).

566   Indochinese-Sundaic faunal transition and phylogeographical divides north of the Isthmus

567   of Kra in Southeast Asian Bulbuls (Aves: Pycnonotidae). Journal of Biogeography

568   43:471–483.

569   Donegan, T., A. Quevedo, J. C. Verhelst, O. Cortés-Herrera, T. Ellery, P. Salaman (2015).

570   Revision of the status of bird species occurring or reported in Colombia, with discussion

571   of BirdLife International's new taxonomy. Conservación Colombiana 23:3–48.

572

573   Donegan, T. M. (2018). What is a species? A new universal method to measure

574   differentiation and assess the taxonomic rank of allopatric populations, using continuous

575   variables. Zootaxa 757:1–67.

576   Donegan, T. M., J. C. Verhelst, T. Ellery, O. Cortés-Herrera, and P. Salaman (2016).

577   Revision of the status of bird species occurring or reported in Colombia 2016 and

578   assessment of BirdLife International's new parrot taxonomy. Conservación Colombiana

579   24:12–36.

580   van Doren, B. M., B. G. Freeman, N. Aristizabal, M. Alvarez-R., J. Pérez-Emán, J., A. M.

581   Cuervo, and G. A. Bravo (2018). Species limits in the Rusty-breasted Antpitta

582   (*Grallaricula ferrugineipectus*) complex. Wilson Journal of Ornithology 130:152–167.

583   Drovetski, S. V., A. B. Reeves, Y. A. Red'kin, I. V. Fadeev, E. A. Koblik, V. N. Sotnikov,

584   and Voelker, G. (2018). Multi-locus reassessment of a striking discord between mtDNA

585   gene trees and taxonomy across two congeneric species complexes. Molecular

586   Phylogenetics and Evolution 120:43–52.

587   van Els, P., and Norambuena, H. V. (2018). A revision of species limits in Neotropical pipits

588   *Anthus* based on multilocus genetic and vocal data. The Ibis 160:158–172.

589   Fernando, S. P., D. E. Irwin, and S. S. Seneviratne (2016). Phenotypic and genetic analysis

590   support distinct species status of the Red-backed Woodpecker (Lesser Sri Lanka Flameback:

591   *Dinopium psarodes*) of Sri Lanka. The Auk 133:497–511.

592   Ferreira, M., A. Aleixo, C. C. Ribas, and M. P. D. Santos (2017). Biogeography of the Neotropical

593   genus *Malacoptila* (Aves: Bucconidae): the influence of the Andean orogeny, Amazonian

594   drainage evolution and palaeoclimate. Journal of Biogeography 44:748–759.

595   Franz, I., D. J. Alvares, and M. Borges-Martins (2020). Species limits in the Tawny-crowned

596   Pygmy-tyrant *Euscarthmus meloryphus* complex (Aves: Passeriformes: Tyrannidae). Zootaxa

597   4809:475–495.

598   Freeman, B. G., and G. A. Montgomery (2017). Using song playback experiments to measure

599   species recognition between geographically isolated populations: a comparison with acoustic

600   trait analyses. The Auk 134:857–870.

601   Friis, G., P. Aleixandre, R. Rodríguez-Estrella, A. G. Navarro-Sigüenza, and B. Milá (2016).

602   Rapid postglacial diversification and long-term stasis within the songbird genus *Junco*:

603   phylogeographic and phylogenomic evidence. Molec. Ecol. 25:6175-6195.

604   Garnett, S. T., and L. Christidis (2017). Taxonomic anarchy hampers conservation. Nature

605   546:25–27.

606   Gwee, C. Y., L. Christidis, J. A. Eaton, J. A. Norman, C. R. Trainor, P. Verbelen, and F. E.

18

607   Rheindt (2017). Bioacoustic and multi-locus DNA data of *Ninox* owls support high incidence of

608   extinction and recolonisation on small, low-lying islands across Wallacea. Molecular

609   Phylogenetics and Evolution 109:246–258.

610   Gwee, C. Y., J. A. Eaton, K. M. Garg, P. Alström, S. van Balen, R. O. Hutchinson, D. M.

611   Prawiradilaga, M. H. Le, and F. E. Rheindt (2019). Cryptic diversity in *Cyornis* (Aves:

612   Muscicapidae) jungle-flycatchers flagged by simple bioacoustics approaches. Zoological

613   Journal of the Linnean Society 186:725–741.

614   Gwee, C. Y., Q. L. Lee, S. P. Mahood, M. H. Le, R. Tizard, K. Elamampai, P. D. Round, and F. E.

615   Rheindt (2020). The interplay of colour and bioacoustic traits in the differentiation of a

616   Southeast Asian songbird complex. Molecular Ecology 30:297–309.

617   Haffer, J. (1992). The history of species concepts and species limits in ornithology. Bulletin

618   British Ornithologists' Club 112A:107–158.

619   Haffer, J. (1997). Species concepts and species limits in ornithology. Pp.11–24 in J. del Hoyo, A.

620   Elliott and J. Sargatal, eds. Handbook of the Birds of the World, 4. Lynx Edicions, Barcelona.

621   Halley, M. R., J. C. Klicka, P. R. Sesink Clee, and J. D. Weckstein (2017). Restoring the species

622   status of *Catharus maculatus* (Aves: Turdidae), a secretive Andean thrush, with a critique of

623   the yardstick approach to species delimitation. Zootaxa 4276:387–404.

624   Hayes, F. E., H. D. Pratt, and C. J. Cianchini (2016). The avifauna of Kosrae, Micronesia:

625   history, status, and taxonomy. Pacific Science 70:91–127.

626   Helbig, A. J., A. G. Knox, D. T. Parkin, D. T., G. Sangster, and M. Collinson (2002).

627   Guidelines for assigning species rank. The Ibis 144:518–525.

628   Hey, J., R. S. Waples, M. L. Arnold, R. K. Butlin, and R. G. Harrison (2003). Understanding

629   and confronting species uncertainty in biology and conservation. Trends in Ecology and

630   Evolution 18:597–603.

631   Hosner, P. A., L. C. Campillo, M. J. Andersen, L. A. Sánchez-González, C. H. Oliveros, R.

632   C. Urriza, and R. G. Moyle (2018). An integrative species delimitation approach reveals

633   finescale endemism and substantial unrecognized avian diversity in the Philippine

634   Archipelago. Conservation Genetics 19:1153–1168.

635   Houle, D., C. Pelabon, G. P. Wagner, and T. F. Hansen (2011). Measurement and meaning in

636   biology. Quarterly Review of Biology 86:1–32.

637   del Hoyo, J., A. Elliott, J. Sargatal, and D. A. Christie (1992–2011). Handbook of the Birds

638   of the World, vols. 1–16. Lynx Edicions, Barcelona.

639   del Hoyo, J., and N. J. Collar (2014–2016). The HBW–BirdLife International Illustrated Checklist

640   of the Birds of the World, vols. 1–2. Lynx Edicions, Barcelona.

641   Huang, Z., F. Tu, and D. Ke (2017). Complete mitochondrial genome of Blue-throated Bee-eater

642   *Merops viridis* (Coraciiformes: Meropidae) with its taxonomic consideration. Pakistan Journal

643   of Zoology 49:79-84.

644   Hung, C.-M., H.-Y. Hung, C.-F. Yeh, Y.-Q. Fu, D. Chen, F. Lei, C.-T. Yao, C.-J. Yao, X.-J. Yang,

645   Y.-T. Lai, and S.-H. Li (2014). Species delimitation in the Chinese bamboo partridge

646   *Bambusicola thoracica* (Phasianidae; Aves). Zoologica Scripta 43:562–575.

647   Hunter, N., M. Mills, and C. Cohen (2019). Elgon Francolin *Scleroptila elgonensis* should be

648   treated as a species distinct from Moorland Francolin *S. psilolaema*. Bulletin African Bird Club

649   26:92–100.

650   Huntley, J. W., and G. Voelker (2016). Cryptic diversity in Afro-tropical lowland forests: the

651   systematics and biogeography of the avian genus *Bleda*. Molecular Phylogenetics and

652   Evolution 99:297–308.

653   Isaac, N. J. B., J. Mallet, and G. M. Mace (2004). Taxonomic inflation: its effect on

654   macroecology and conservation. Trends in Ecology and Evolution 19:464–469.

655   Isler, M. L., and B. M. Whitney (2017). Species limits in the genus *Thamnistes* (Aves:

656   Passeriformes: Thamnophilidae): an evaluation based on vocalizations. Zootaxa 4291:192–200.

657   Isler, M. L., P. R. Isler, and B. M. Whitney (1998). Use of vocalizations to establish species limits

658   in antbirds (Passeriformes: Thamnophilidae). The Auk 115:577–590.

659   Johansson, U. S., M. Irestedt, Y. Cu, and P. G. P. Ericson (2018). Phylogenetic relationships of

660   rollers (Coraciidae) based on complete mitochondrial genomes and fifteen nuclear genes.

661   Molecular Phylogenetics and Evolution 126:17–22.

662   Johnson, N. K., J. V. Remsen, and C. Cicero (1999). Resolution of the debate over species

663   concepts in ornithology: a new comprehensive biologic species concept. Pp.1470-1482 in Proc.

664   22 Internatnational Ornithological Congress, BirdLife South Africa, Durban.

665   Jønsson, K. A., K. Delhey, G. Sangster, P. G. P. Ericson, and M. Irestedt (2016). The evolution of

666   mimicry of friarbirds by orioles (Aves: Passeriformes) in Australo-Pacific archipelagos. Proc.

667   R. Soc. B 283:20160409.

668   Jønsson, K. A., M. P. K. Blom, P. Z. Marki, L. Joseph, G. Sangster, P. G. P. Ericsson, and M.

669   Irestedt (2019a). Complete subspecies-level phylogeny of the Oriolidae (Aves: Passeriformes):

670   out of Australasia and return. Molecular Phylogenetics and Evolution 137:200–209.

671   Jønsson, K. A., A. H. Reeve, M. P. K. Blom, M. Irestedt, and P. Z. Marki (2019b). Unrecognised

672   (species) diversity in New Guinean passerine birds. The Emu 119:233–241.

673   Kearns, A. M., M. Restani, I. Szabo, A. Schrøder-Nielsen, J. A. Kim, H. M. Richardson, J. M.

674   Marzluff, R. C. Fleischer, A. Johnsen, and K. E. Omland (2018). Genomic evidence of

675   speciation reversal in ravens. Nature Communications 9:906.

676 Kirchman, J. J., C. C. Witt, J. A. McGuire, and G. R. Graves (2010). DNA from a 100-year-old

677   holotype confirms the validity of a potentially extinct hummingbird species. Biology Letters

678   6:112–115.

679 Kryukov, A. P., L. N. Spiridonova, S. Mori, V. Yu. Arkhipov, Y. A. Red'kin, O. A. Goroshko, E.

680   G. Lobkov, and E. Haring (2017). Deep phylogeographic breaks in Magpie *Pica pica* across the

681   Holarctic: concordance with bioacoustics and phenotypes. Zoological Science 34:185–200.

682 Lees, A. L., and S. L. Pimm (2015). Species, extinct before we know them? Current Biology

683   25:R177–R180.

684 Manthey, J. D., M. Geiger, and R. G. Moyle (2017). Relationships of morphological groups in the

685   northern flicker superspecies complex (*Colaptes auratus* & *C. chrysoides*). Systematics and

686   Biodiversity 15:183–191.

687 Martens, J., and N. Bahr (2016). Dokumentation neuer Vogel-Taxa, 10—Bericht für 2014.

688   Vogelwarte 54:195–230.

689 Martens, J., Y. Sun, and M. Päckert (2008). Intraspecific differentiation of Sino-Himalayan

690   bushdwelling *Phylloscopus* leaf warblers, with description of two new taxa (*P. fuscatus*, *P.*

691   *fuligiventer*, *P. affinis*, *P. armandii*, *P. subaffinis*). Vertebrate Zoology 58:233–265.

692 Mason, N. A., A. Olvera-Vital, I. J. Lovette, and A. G. Navarro-Sigüenza (2018). Hidden

693   endemism, deep polyphyly, and repeated dispersal across the Isthmus of Tehuantepec:

694   diversification of the White-collared Seedeater complex (Thraupidae: *Sporophila torqueola*).

695   Ecology and Evolution 8:1867–1881.

696 Mayr, E. (1946). The number of species of birds. The Auk 63:64-69.

697 Mayr, E. (1969). Principles of Systematic Zoology. McGraw-Hill, New York.

698 Meiri, S., and G. M. Mace (2007). New taxonomy and the origin of species. PLoS Biology 5:e194.

699 Miller, M. J., E. Bermingham, B. L. Turner, S. E. Lipshutz, J. C. Touchon, A. B. Johnson, and K.

700   Winker (2021). Demographics of foraging ecology explain genetic diversification in

701   Neotropical bird species. Ecology Letters In press. https://doi.org/10.1111/ele.13674.

702 Monk, J. F., ed. (1992). Avian systematics and taxonomy. Bulletin British Ornithologists' Club

703   112A.

704 Moyle, R. G., J. D. Manthey, P. A. Hosner, M. Rahman, M. Lakim, and F. H. Sheldon (2017). A

705   genome-wide assessment of stages of elevational parapatry in Bornean passerine birds reveals

706   no introgression: implications for processes and patterns of speciation. PeerJ 18:e3335.

707 Päckert, M., J. Martens, M. Wink, A. Feigl, and D. T. Tietze (2012). Molecular phylogeny of Old

708   World swifts (Aves: Apodiformes, Apodidae, *Apus* and *Tachymarptis*) based on mitochondrial

709   and nuclear markers. Molecular Phylogenetics and Evolution 63:606–616.

710   Padial, J. M., A. Miralles, I. De la Riva, and M. Vences (2010). The integrative future of

711   taxonomy. Frontiers in Zoology 7:1–14.

712   Pearson, D. J., and D. A. Turner (2017). A taxonomic review of the genus *Zosterops* in East

713   Africa, with a revised list of species occurring in Kenya, Uganda and Tanzania. Scopus 37:1–

714   13.

715   Pedersen, M. P., M. Irestedt, L. Joseph, C. Rahbek, and K. A. Jønsson (2018). Phylogeography of

716   a 'great speciator' (Aves: *Edolisoma tenuirostre*) reveals complex dispersal and diversification

717   dynamics across the Indo-Pacific. Journal of Biogeography 45:826–837.

718   Pellegrino, I., M. Cucco, J. A. Harvey, F. Liberatore, M. Pavia, G. Voelker, and G. Boano (2017).

719   So similar and yet so different: taxonomic status of Pallid Swift *Apus pallidus* and Common

720   Swift *Apus apus*. Bird Study 64:344–352.

721   Pérez-Emán, J. L., J. P. Ferreira, N. Gutiérrez-Pinto, A. M. Cuervo, L. N. Céspedes, C. C. Witt,

722   and C. D. Cadena (2018). An extinct hummingbird species that never was: a cautionary tale

723   about sampling issues in molecular phylogenetics. Zootaxa 4442:491–497.

724   Peters, J. L., and successors (1931–1986). Check-list of Birds of the World, 1–15. Harvard

725   University Press/Museum of Comparative Zoology, Cambridge, Mass.

726   Remsen Jr, J. V. (2016). A 'rapid assessment program' for assigning species rank? Journal of

727   Field Ornithology 87:110–115.

728   Rheindt, F. E., and S. V. Edwards (2011). Genetic introgression: An integral but neglected

729   component of speciation in birds. The Auk 128:620–632.

730   Sadanandan, K. R., C. Küpper, G. W. Low, C.-T. Yao, Y. Li, T. Xu, F. E. Rheindt, and S. Wu

731   (2019). Population divergence and gene flow in two East Asian shorebirds on the verge of

732   speciation. Scientific Reports 9:8546.

733   Sadanandan, K. R., G. W. Low, S. Sridharan, C. Y. Gwee, E. Y. X. Ng, P. Yuda, D. M.

734   Prawiradilaga, J. G. H. Lee, A. Tritto, and F. E. Rheindt (2020). The conservation value of

735   admixed phenotypes in a critically endangered species complex. Scientific Reports 10:15549.

736   Sandoval, L., P. P. Bitton, A. D. Demko, S. M. Doucet, and D. J. Mennill (2017). Phenotypic

737   variation and vocal divergence reveals a species complex in White-eared Ground-sparrows

738   (Cabanis) (Aves: Passerellidae). Zootaxa 4291:155–170.

739   Sangster, G. (2018). Integrative taxonomy of birds: the nature and delimitation of species. Pp.9-38

740   in D. T. Tietze, ed. Bird Species: How They Arise, Modify and Vanish. Springer Nature, Cham,

741   Switzerland.

742   Shakya, S. B., J. Fuchs, J.-M. Pons, and F. H. Sheldon (2017). Tapping the woodpecker tree for

743   evolutionary insight. Molecular Phylogenetics and Evolution 116:182–191.

744   Sharpe, R. B. (1899–1909). A Hand-list of the Genera and Species of Birds. Trustees of the British

745   Museum, London.

746   Sibley, C. G., and Monroe, B. L. (1990). Distribution and Taxonomy of Birds of the World. Yale

747   University Press, New Haven, CT, USA.

748   Tobias, J. A., N. Seddon, C. N. Spottiswoode, J. D. Pilgrim, L. D. C. Fishpool, and N. J. Collar

749   (2010). Quantitative criteria for species delimitation. The Ibis 152:724-746.

750   Tobias, J. A., J. Ottenburghs, and A. Pigot (2020). Avian diversity: speciation, macroevolution

751   and ecological function. Annual Reviews of Ecology, Evolution and Systematics 51:533–560.

752   Toews, D. P. L., S. A. Taylor, R. Vallender, A. Brelsford, B. G. Butcher, P. W. Messer, and I. J.

753   Lovette (2016). Plumage genes and little else distinguish the genomes of hybridizing warblers.

754   Current Biology 26:2313–2318.

755   Töpfer, T., L. Podsiadlowski, and K. Gedeon (2014). Rediscovery of the Black-fronted Francolin

756   *Pternistis* (*castaneicollis*) *atrifrons* (Conover, 1930) (Aves: Galliformes: Phasianidae) with

757   notes on biology, taxonomy and conservation. Vertebrate Zoology 64:261–271.

758   Turner, D. A., F. Dowsett-Lemaire, and R. J. Turner (2020). Further support for the specific

759   distinctness of Elgon Francolin *Scleroptila elgonensis*. Bulletin African Bird Club 27:80–83.

760   Wang, X., P. Que, G. Heckel, J. Hu, X. Zhang, C.-Y. Chiang, N. Zhang, Q. Huang, S. Liu, J.

761   Martinez, E. Pagani-Núñez, C. Dingle, Y. Y. Leung, T. Székely, Z. Zhang, and Y. Liu (2019).

762   Genetic, phenotypic and ecological differentiation suggests incipient speciation in two

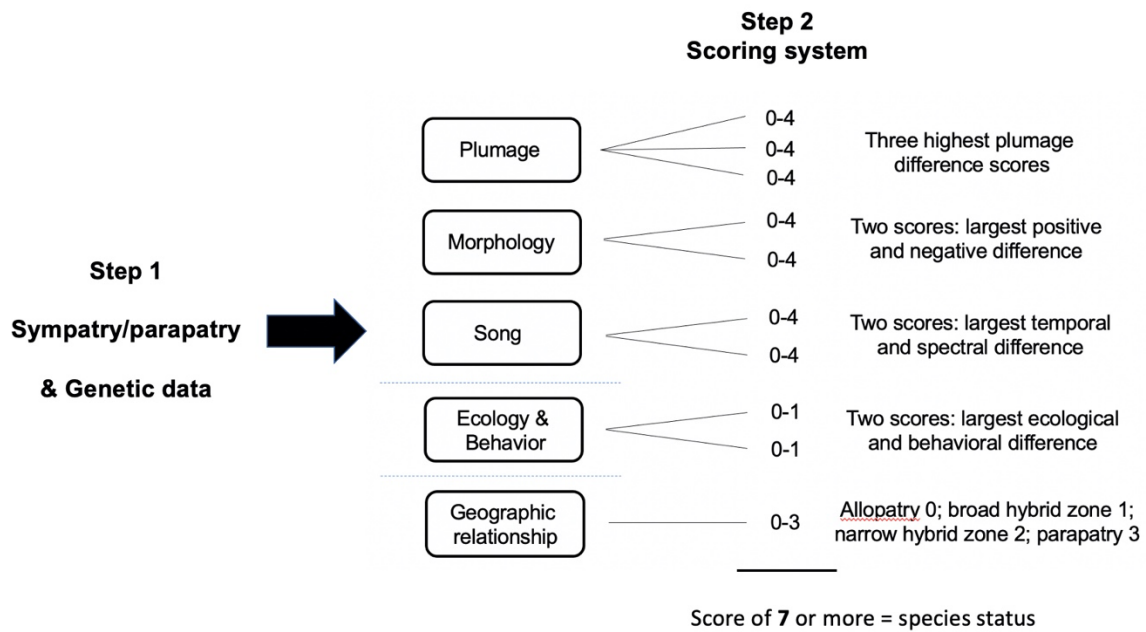763   *Charadrius* plovers along the Chinese coast. BMC Evolutionary Biology 19:135.

764   Winker, K. (2009) Reuniting phenotype and genotype in biodiversity research. BioScience

765   59:657–665.

766   Winker, K. (2016). An examination of species limits in the *Aulacorhynchus* '*prasinus*' toucanet

767   complex (Aves: Ramphastidae). PeerJ. 4:e2381.

768   Winker, K., D. A. Rocque, T. M. Braile, and C. L. Pruett (2007). Vainly beating the air: species-

769   concept debates need not impede progress in science or conservation. Ornithological

770   Monographs 63:30–44.

771   Jiao, X., and Z. Yang (2021). Defining species when there is gene flow. *Systematic Biology* 70: 108–

772   119.

773   Zhang, D., L. Tang, Y. Cheng, Y. Hao, Y. Xiong, G. Song, Y. Qu, F. E. Rheindt, P. Alström, C.

774   Jia, and F. Lei (2019). 'Ghost introgression' as a cause of deep mitochondrial divergence in a

775   bird species complex. Molecular Biology and Evolution 36:2375–2386.

Step 2
Scoring system

Plumage	0-4 / 0-4 / 0-4	Three highest plumage difference scores
Morphology	0-4 / 0-4	Two scores: largest positive and negative difference
Song	0-4 / 0-4	Two scores: largest temporal and spectral difference
Ecology & Behavior	0-1 / 0-1	Two scores: largest ecological and behavioral difference
Geographic relationship	0-3	Allopatry 0; broad hybrid zone 1; narrow hybrid zone 2; parapatry 3

Step 1

Sympatry/parapatry

& Genetic data

Score of **7** or more = species status

776
777
778

779 **FIGURE 1.** Diagram of two-step application of taxonomic criteria proposed by Tobias et al.

780 (2010). Step 1 gives priority to genetic data in cases where taxa are in confirmed contact (e.g.

781 sympatric or parapatric). If molecular evidence indicates substantial divergence across

782 multiple loci, or low levels of gene flow between such taxa, they are treated as species

783 regardless of their level of phenotypic divergence. If molecular evidence is inconclusive, or

784 the taxa are allopatric, their level of divergence is scored in Step 2. No individual character

785 can score more than 4, with 4 reflecting 'exceptional divergence' in a major character

786 (particularly those involved in reproductive isolation, e.g. mating display, song or ornament).

787 To limit multiple counting of correlated traits, scores are capped to two orthogonal

788 morphometric and vocal characters, defined by thresholds in effect size (Cohen's d)

789 calculated from measurements taken from >10 individuals/songs.

790
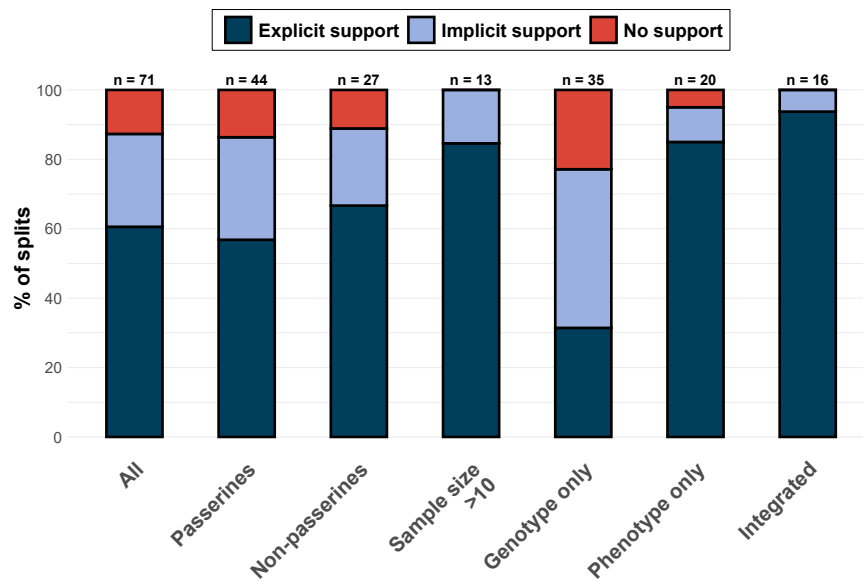
791

792

793   **FIGURE 2.** Degree of concordance (%) between taxonomic splits made by application of the

794   criteria of Tobias *et al.* (2010) and subsequent, independent taxonomic research. Levels of

795   explicit and implicit support are high for all splits that have been subsequently assessed (All).

796   In addition, despite the criteria being calibrated with a sample heavily biased toward

797   passerines, support was high for splits made in both passerines and non-passerines,

798   separately. In cases where the criteria were applied using >10 individuals for morphometric

799   analyses, in line with 'High' robustness (see Methods) and following recommendations of

800   Tobias *et al.* (2010), independent support for the split was 100%. Support was relatively low

801   in studies focusing only on molecular data ('Genotype only'), much higher in studies

802   focusing only on phenotypic data, and 100% in studies combining a range of data sources

803   (Integrated). Sample sizes (number of splits) are given above the bars.

TABLE 1. List of splits made using scoring criteria which have subsequently been independently assessed. Robustness of criteria application is scored largely in respect of sample size (High: >10 individuals sampled) and whether data on vocalisations were available (see Methods). The method(s) used in the subsequent independent assessment are scored: A = genotype, B = phenotype, C = vocalizations, D = other. Thus, a study marked 'A,C' used a combination of genetic and vocal data. Degree of support indicates the extent to which the subsequent independent assessment supported the split made using the scoring criteria (see Methods).

Parent Taxon	Split	Robustness of criteria application	Method in subsequent assessment	Degree of support	Source
Acridotheres melanopterus	Acridotheres tricolor	Medium	A	None[1]	Sadanandan et al. 2020
Acridotheres melanopterus	Acridotheres tertius	Medium	A	None[1]	Sadanandan et al. 2020
Alcedo cyanopectus	Ceyx nigrirostris	Low	A	Implicit	Andersen et al. 2018
Actenoides monachus	Actenoides capucinus	Medium	A	None	Andersen et al. 2018
Amazona autumnalis	Amazona lilacina	Medium	B	Explicit	Donegan et al. 2016
Amazona festiva	Amazona bodini	High	B	Explicit	Donegan et al. 2016
Anthus lutescens	Anthus peruvianus	Medium	A,C	Explicit	van Els et al. 2018
Arremon taciturnus	Arremon axillaris	Medium	B,C	Explicit	Buainain et al. 2017
Aulacorhynchus prasinus	A. cyanolaemus	Medium	B,D	None[2]	Winker 2016
Bambusicola thoracica	Bambusicola sonorivox	High	A,C	Explicit[3]	Hung et al. 2014
Basileuterus culicivorus	Basileuterus cabanisi	High	C	Explicit	Freeman and Montgomery 2017
Basileuterus luteoviridis	Myiothlypis striaticeps	Low	C	Explicit	Freeman and Montgomery 2017

Bleda notatus	*Bleda ugandae*	Medium	A	Explicit[3]	Huntley and Voelker 2016
Bowdleria punctata	*Poodytes caudatus*	Medium	A	Implicit	Alström *et al.* 2018a
Brachypteryx montana	*Brachypteryx erythrogyna*	Medium	A	Implicit	Kyriazis *et al.* 2018
Brachypteryx montana	*Brachypteryx poliogyna*	Medium	A	Implicit	Kyriazis *et al.* 2018
Brachypteryx montana	*Brachypteryx cruralis*	Low	A,B,C,D	Explicit	Alström *et al.* 2018b
Brachypteryx montana	*Brachypteryx goodfellowi*	Medium	A,B,C,D	Explicit	Alström *et al.* 2018b
Brachypteryx montana	*Brachypteryx sinensis*	Medium	A,B,C,D	Explicit	Alström *et al.* 2018b
Ceyx melanurus	*Ceyx mindanensis*	Low	A	Implicit	Andersen *et al.* 2018b
Charadrius alexandrinus	*Charadrius dealbatus*	High	A,B,D	Explicit	Sadanandan *et al.* 2019, Wang *et al.* 2019
Charadrius obscurus	*Charadrius aquilonius*	Medium	A	None[3]	Barth *et al.* 2013
Cittura cyanotis	*Cittura sanghirensis*	High	A	Implicit	Andersen *et al.* 2018
Colaptes auratus	*Colaptes mexicanoides*	Medium	A	Implicit	Manthey *et al.* 2017
Coracias benghalensis	*Coracias affinis*	Medium	A	Explicit	Johansson *et al.* 2018
Cyornis tickelliae	*Cyornis sumatranus*	Low	C	Explicit	Gwee *et al.* 2019
Dinopium benghalense	*Dinopium psarodes*	Medium	A	Explicit	Fernando *et al.* 2016
Edolisoma tenuirostre	*Edolisoma grayi*	Medium	A	None	Pedersen *et al.* 2018
Edolisoma tenuirostre	*Edolisoma obiense*	Medium	A	None	Pedersen *et al.* 2018
Euscarthmus meloryphus	*Euscarthmus fulviceps*	Medium	B,C	Explicit	Franz *et al.* 2020
Forpus xanthopterygius	*Forpus spengeli*	Medium	B	Explicit[4]	Bocalini and Silveira 2015
Francolinus castaneicollis	*Pternistis atrifrons*	Medium	A,B,C	Explicit[3]	Töpfer *et al.* 2014

Francolinus psilolaemus	*Scleroptila elgonensis*	Medium	A,B,C	Explicit	Hunter *et al.* 2019, Turner *et al.* 2020
Goura scheepmakeri	*Goura sclaterii*	Medium	A	Explicit	Bruxaux *et al.* 2018
Gracupica contra	*Gracupica jalla*	High	A	Explicit	Baveja *et al.* 2020
Grallaria quitensis	*Grallaria alticola*	Low	C	Explicit	Freeman and Montgomery 2017
Grallaricula ferrugineipectus	*Grallaricula leymebambae*	Low	A,B,C	Explicit	van Doren *et al.* 2018
Halcyon smyrnensis	*Halcyon gularis*	High	A	Implicit	Andersen *et al.* 2018
Junco phaeonotus	*Junco bairdi*	Medium	A,B	Implicit[3]	Friis *et al.* 2016
Lacedo pulchella	*Lacedo melanops*	Medium	A	Implicit	Andersen *et al.* 2018
Macronous flavicollis	*Mixornis prillwitzi*	Medium	C	Implicit	Cros and Rheindt 2017
Malacoptila striata	*Malacoptila minor*	Medium	A	Explicit	Ferreira *et al.* 2017
Melozone leucotis	*Melozone occipitalis*	Medium	B,C	Explicit	Sandoval *et al.* 2017
Merops viridis	*Merops americanus*	Medium	A	Explicit	Huang *et al.* 2017
Mulleripicus funebris	*Mulleripicus fuliginosus*	High	A	Explicit	Shakya *et al.* 2017
Ninox squamipila	*Ninox hantu*	High	A,C	Explicit	Gwee *et al.* 2017
Oriolus melanotis	*Oriolus finschi*	Low	A	Implicit[3]	Jønsson *et al.* 2016
Oriolus cruentus	*Oriolus consanguineus*	Medium	A	Implicit	Jønsson *et al.* 2019a
Paramythia montium	*Paramythia olivacea*	Medium	A	Implicit	Jønsson *et al.* 2019b
Pica pica	*Pica asirensis*	Low	A	Explicit	Kryukov *et al.* 2017
Pica pica	*Pica mauritanica*	Medium	A,C	Explicit	Kryukov *et al.* 2017
Pomatorhinus erythrocnemis	*Erythrogenys gravivox*	High	A	Explicit	Dai *et al.* 2019

Pomatorhinus erythrocnemis	*Erythrogenys swinhoei*	High	A	Explicit	Dai *et al.* 2019
Psittacara wagleri	*Psittacara frontatus*	High	B	Explicit	Donegan *et al.* 2016
Ptilinopus porphyraceus	*Ptilinopus hernsheimi*	Medium	C	Explicit	Hayes *et al.* 2016
Pycnonotus flavescens	*Pycnonotus leucops*	Medium	A	Implicit[3]	Dejtaradol *et al.* 2016
Pyrocephalus rubinus	*Pyrocephalus dubius*	Medium	A,B,C	Explicit[3]	Carmi *et al.* 2016
Pyrocephalus rubinus	*Pyrocephalus nanus*	Medium	A,B,C	Explicit[3]	Carmi *et al.* 2016
Pyrrhura melanura	*Pyrrhura pacifica*	Low	B	Explicit	Donegan *et al.* 2016
Sittasomus griseicapillus	*Sittasomus griseus*	Low	C	Explicit	Freeman and Montgomery 2017
Sporophila torqueola	*Sporophila morelleti*	Medium	A,B	Explicit	Mason *et al.* 2018
Stephanoxis lalandi	*Stephanoxis loddigesii*	High	B	Explicit[3]	Cavarzere *et al.* 2014
Thamnistes anabatinus	*Thamnistes aequatorialis*	Low	C	Implicit[5]	Isler *et al.* 2017
Thryothorus euophrys	*Pheugopedius schulenbergi*	Low	C	Explicit	Freeman and Montgomery 2017
Turdinus crispifrons	*Gypsophila calcicola*	Low	A,B,C	Explicit[6]	Gwee *et al.* 2020
Zosterops cinereus	*Zosterops ponapensis*	Medium	B,C	Explicit[3]	Hayes *et al.* 2016
Zosterops poliogastrus	*Zosterops eurycricotus*	Low	A[7]	Implicit	Pearson and Turner 2017
Zosterops poliogastrus	*Zosterops mbuluensis*	Low	A[7]	Implicit	Pearson and Turner 2017
Zosterops poliogastrus	*Zosterops winifredae*	Low	A[7]	Implicit	Pearson and Turner 2017
Zosterops poliogastrus	*Zosterops kaffensis*	Low	A[7]	None	Pearson and Turner 2017
Zosterops poliogastrus	*Zosterops kulalensis*	Low	A[7]	None	Pearson and Turner 2017

809   1.   Authors suggest that the forms should be treated as separate conservation units

810   2.   Winker (2016) otherwise concurs with all species limits as defined under the criteria in the *A. prasinus* group

811   3. Paper published concurrently or immediately prior to Checklist but only seen subsequently

812   4. Split supported by Donegan *et al*. (2016), although they speculated that *spengelii* might be a subspecies of *F. passerinus*

813   5. Isler *et al.* selected only *T. rufescens* to split; the criteria split *rufescens*, *aequatorialis* and *gularis*, Isler *et al.* conceded *aequatorialis* is vocally distinct

814   from the four trans-Andean taxa

815   6. Gwee *et al*. 2020 also split *annamensis* but this was not scored using the criteria owing to lack of material

816   7. Authors reviewed molecular evidence from earlier studies which had made no taxonomic recommendations

817