O p i n i o n   p a p e r

urn:lsid:zoobank.org:pub:A6657A24-5047-4262-BAB5-BB0FD4991129

Measuring and explaining disagreement in bird taxonomy

Stijn CONIX [1,*], Vincent CUYPERS [2] & Charles H. PENCE [3]

[1] Institut supérieur de philosophie, Université catholique de Louvain, Louvain-La-Neuve, Belgium.
[2] Research group zoology: biodiversity and toxicology, Centre for environmental sciences,
Hasselt university, Diepenbeek, Belgium.
[2] Centre for logic and philosophy of science, KU Leuven, Leuven, Belgium.
[3] Institut supérieur de philosophie, Université catholique de Louvain, Louvain-La-Neuve, Belgium.

[*] Corresponding author: stijn.conix@uclouvain.be
[2] Email: Vincent.cuypers@uhasselt.be
[3] Email: charles.pence@uclouvain.be

[1] urn:lsid:zoobank.org:author:48745EBD-207C-4D70-82AE-2AEF0E135145
[2] urn:lsid:zoobank.org:author:010A1B1F-C2F8-4098-BCBC-6BCD405A5016
[3] urn:lsid:zoobank.org:author:001C47BE-9729-4779-93BA-6FEEBB12FD3B

Abstract. Species lists play an important role in biology and practical domains like conservation, legislation, biosecurity and trade regulation. However, their effective use by non-specialist scientific and societal users is sometimes hindered by disagreements between competing lists. While it is well-known that such disagreements exist, it remains unclear how prevalent they are, what their nature is, and what causes them. In this study, we argue that these questions should be investigated using methods based on taxon concept rather than methods based on Linnaean names, and use such a concept-based method to quantify disagreement about bird classification and investigate its relation to research effort. We found that there was disagreement about 38% of all groups of birds recognized as a species, more than three times as much as indicated by previous measures. Disagreement about the delimitation of bird groups was the most common kind of conflict, outnumbering disagreement about nomenclature and disagreement about rank. While high levels of conflict about rank were associated with lower levels of research effort, this was not the case for conflict about the delimitation of bird groups. This suggests that taxonomic disagreement cannot be resolved simply by increasing research effort.

Keywords. Taxonomic disagreement, taxonomic disorder, species lists, bird lists, taxonomic conflict.

Introduction

Species lists play an important role in a range of societal domains. They underpin scientific research in most disciplines of biology, and are a crucial tool for biodiversity conservation, trade regulation and biosecurity. Stakeholders in these domains look to taxonomists to provide these lists, and expect them to be clear and accurate (Garnett *et al.* 2020; Conix *et al.* 2021; Thiele *et al.* 2021). However, for many

branches of the Tree of Life no good and recent taxonomies are available due to a lack of research, and for several other branches there is enduring disagreement about which taxonomy to adopt (Isaac *et al.* 2004; Garnett & Christidis 2017). Such disagreement leads to the publication of competing taxonomic treatments for taxa, making it hard for users to know which classification to use. In practice, this leads to confusion, difficulties in communication and the exchange of data, and in some cases the use of inadequate classifications in relation to specific scientific or policy-related objectives (Thomson *et al.* 2021).

Various solutions have been proposed to deal with this problem. For example, some propose forms of *taxonomic governance*, in which an authoritative body curates one single species list that is set as reference standard (Garnett *et al.* 2020; Lien *et al.* 2021). Others propose a system of taxonomic alignment, in which a data tool is constructed that contains all available taxonomies and maps the relations between them, so that they can coexist and can be used alongside each other in an orderly manner (Franz *et al.* 2016; Sterner *et al.* 2020; Cuypers *et al.* 2022). However, solving the problem of taxonomic disagreement, or at least mitigating its consequences, requires in the first place a full understanding of which taxa taxonomists disagree about, what exactly they disagree about, and what causes them to disagree.

While large-scale empirical research on this matter is still scarce, attention for it is increasing. In particular, two recent studies, by McClure *et al.* (2020) and Neate-Clegg *et al.* (2021), trace differences between four widely used global lists of bird species to chart taxonomic disagreement with regard to birds. These four lists are the Handbook of the Birds of the World and Birdlife International Digital Checklist of the Bird of the World (BL hereafter, Birdlife International 2020), the Clements Checklist of Birds of the World (CLEM hereafter, Clements *et al.* 2019), The Howard and Moore Complete Checklist of the Birds of the World (HM hereafter, Dickinson & Christidis 2014) and the IOC World Bird List (IOC hereafter, Gill *et al.* 2021).

Both studies report substantial levels of disagreement between these four species lists and explain it with three main hypotheses. First, they suggest that taxonomic disagreement can be caused by a lack of research effort: the less a taxon has been studied, it is argued, the higher the chance that taxonomists disagree about its classification. Second, these studies point to ecological and biogeographical explanations. More particularly, they claim that more taxonomic disagreement can be expected for taxa showing recent diversification. The idea is that species limits in rapidly and recently diversifying taxa will be harder to interpret because they are likely to contain many groups in the grey area between what are typically treated as species and as subspecies. Finally, these studies also point to the use of different definitions of species – so-called *species concepts*, such as the Biological Species Concept (BSC) – as an important source of disagreement between lists.

While these are all plausible hypotheses in theory, they are far from conclusively demonstrated, as both these recent studies have their limitations. On the one hand, they are limited in the way they measure disagreement. On the other hand, more fine-grained evidence is needed to confirm or disconfirm any of the proposed hypotheses.

With regard to limitations in the measurement of disagreement, McClure *et al.* (2020) only focus on raptors, which are generally more charismatic and more threatened than other birds, and constitute only around 5% of all bird species. It is unclear whether their results can be generalized to other birds. Neate-Clegg *et al.* (2021), on the other hand, take into account all birds, but use a method of measuring disagreement that we think has important disadvantages. While McClure *et al.* use a method based on *taxon concepts*, Neate-Clegg *et al.* use *taxon names* to measure disagreement.

A taxon concept – not to be confused with *species concepts* such as the BSC – represents any group of organisms asserted to be a taxon by some authority, be it as a species, subspecies, or taxon at another level (Berendsohn 1995; Lepage *et al.* 2014). Ideally, such concepts are somehow given a unique identifier (Fig. 1). The advantage of working with such uniquely identifiable taxon concepts, alongside traditional taxonomic names and descriptions, is that concepts unambiguously identify the same group of organisms at all times regardless of the taxonomic status the group in question is given in practice. That makes them more stable than taxonomic names. Names only refer unambiguously to type specimens, and the circumscribed group of organisms they represent often changes over time (Franz *et al.* 2008; Lepage *et al.* 2014). We argue below that, for that reason, measures of taxonomic disagreement that are based on taxonomic names are likely to underestimate conflict and misrepresent its causes.

The existing studies are also limited in how they explain disagreement. Neate-Clegg *et al.* relate disagreement to a range of characteristics of the birds in question (such as body mass, migration behavior, island dependency, and forest dependency), and claim that these relations show that taxonomic disagreement increases with a lack of research effort and with higher levels of (recent) divergence. For example, they hypothesize that taxonomic disagreement decreases with research effort as there is fewer disagreement about birds living in open habitats than about birds living in forests. However, the association between disagreement and these habitats is insufficient to support this hypothesis. After all, it is both conceivable that birds in open habitats show different speciation patterns than forest birds and that birds living in open habitats are more or less studied than forest birds – there are accounts of *forest bias* in biological research (Trimble & van Aarde 2012). In other words, the relations between disagreement and the traits that Neate-Clegg *et al.* investigated are most likely mediated by both research effort and recent divergence. Without a clear causal model that incorporates both these factors, statistical models showing associations between traits such as forest-dependency, migration behavior or habitat on the one hand and taxonomic disagreement on the other cannot separate the roles of research effort and recent divergence in causing disagreement. However valuable the relations observed by Neate-Clegg *et al.* are – they point to hotspots of disagreement meriting particular attention – different analyses are needed to understand and separate the influence of the various drivers of disagreement.
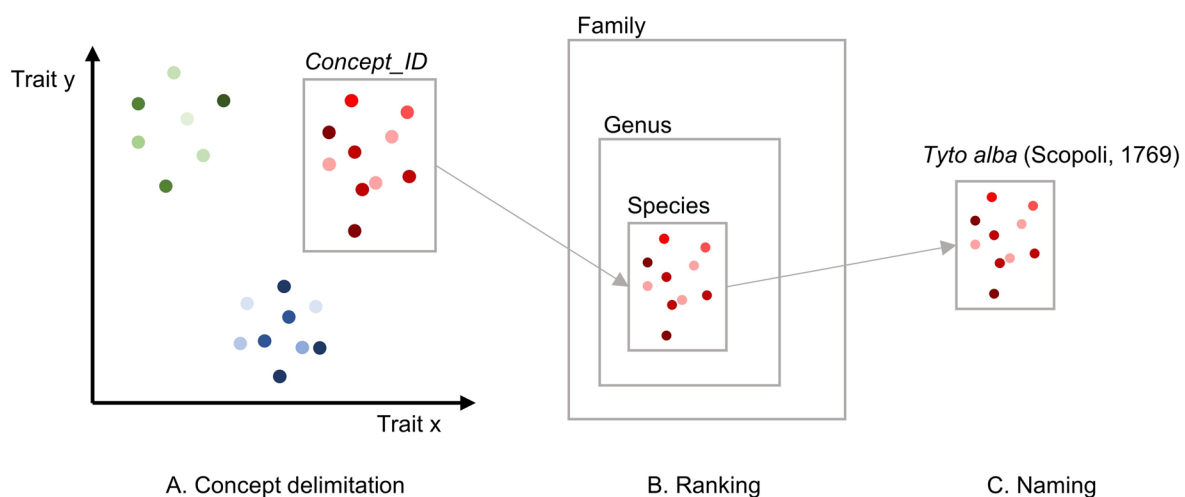

Fig. 1. Schematic representation of the taxonomic research process from the viewpoint of taxon concepts. Disagreements can occur at any step. **A**. The taxonomist groups a number of organisms in a concept, that is given a unique identifier. This grouping can be done using any taxonomic procedure, based on morphological characters, molecular characters, geography, etc. **B**. The concept is introduced in the Linnean taxonomic system and given a rank. **C**. The ranked taxon is given a scientific name.

In this article, we aim to build on both studies of disagreement in bird classification and – in part – overcome their limitations. We do this by using a method based on taxon concepts, in line with the method used by McClure *et al.*, to measure and explore disagreement about all bird species. Moreover, in contrast to McClure *et al.*, our method makes a distinction between three different kinds of conflict in bird taxonomy: disagreement about the taxonomic concepts lists recognize (concept conflict), about the ranks (rank conflict) they give to the concepts they recognize, and about nomenclature (nomenclatural conflict, see next section). Then, we exploratorily associate these different kinds of conflict with three direct proxies of research effort. Using these proxies, we can get a first clear indication on whether trends in research effort can at least partly explain trends in disagreement – alongside or in contrast with patterns of speciation.

On measuring disagreement: concepts or names?

Before delving into the details of our study, it is worth further clarifying the notion of taxonomic disagreement, what taxonomic disagreement is about, and how it ideally ought to be quantified.

Consider the case of the *Tyto alba* (barn owl) complex (Strigiformes: Tytonidae), a controversial case also cited by McClure *et al.* (2020). All four lists (in the versions used for this study) recognize the scientific name *T. alba* as a species-level taxon, but they all disagree about which traditionally recognized subspecific groups are included in the taxon with that name (Fig. 2) (Uva *et al.* 2018). BL uses the name *T. alba* in the most inclusive way, recognizing it to contain the conventional subgroups 'alba' (old world),
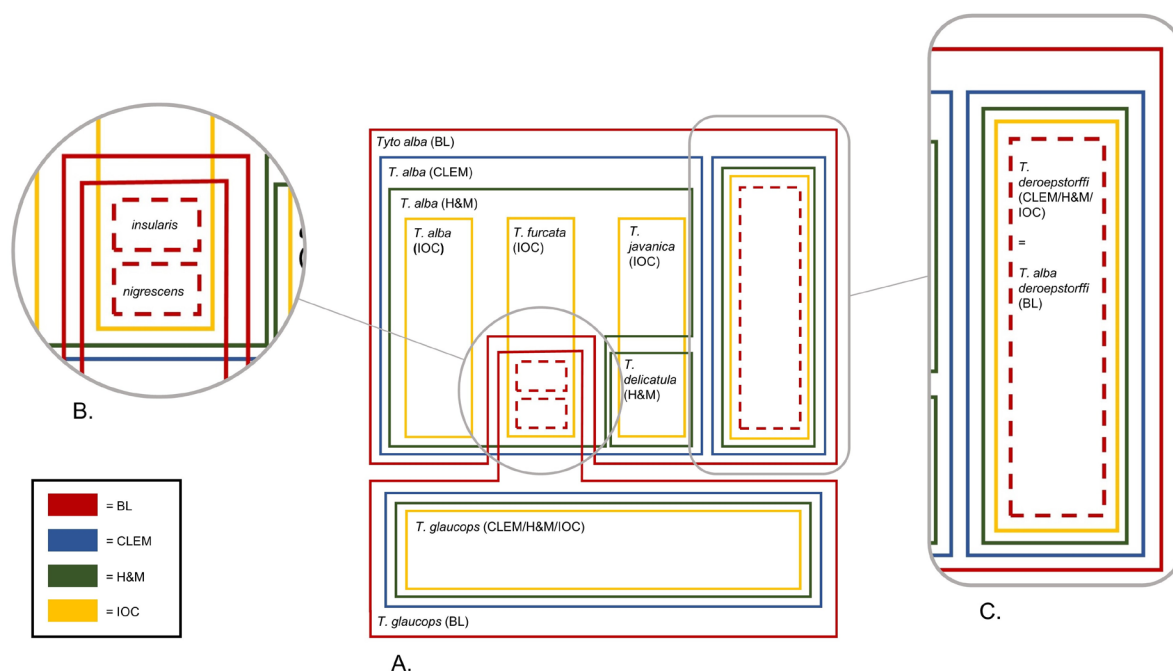

Fig. 2. Schematic representation of some of the taxon concepts recognized by the different lists within the *Tyto alba* complex. Full lines represent concepts with species rank; dashed lines concept at subspecies rank. **A**. General overview. Not all concepts recognized as subspecies are shown. **B**. Example of concept conflict: the subgroups '*insularis*' and '*nigrescens*' belong in *T. alba* according to CLEM and HM; in *T. furcata* according to IOC (because they split *T. furcata* from *T. alba* sensu lato); but in *T. glaucops* according to BL, a taxon that all lists recognize at species-rank. **C**. Example of rank conflict. All lists recognize the same taxon concept, but CLEM, HM and IOC recognize it at species-level (*T. deroepstorffi*), while BL recognizes it at subspecies level (*T. alba deroepstorffi*).

'furcata' (new world), 'javanica' (eastern) and 'deroepstorffi' (Andaman islands). For CLEM, it contains 'alba', 'furcata', 'javanica' but not 'deroepstorffi', which is recognized as a separate species. Moreover, CLEM includes under *T. alba* two Caribbean lover level subgroups, 'insularis' and 'nigrescens', that BL assigns to another species, *T. glaucops*. HM more or less follows CLEM, except in that it excludes an Australian taxon 'delicatula', part of the 'javanica' subgroup, from *T. alba,* recognizing it as a separate species (*T. delicatula*). The IOC list in its turn recognizes *T. alba*, *T. furcata*, *T. javanica*, and *T. deroepstorffi* all as separate species, but lumps *T. delicatula* (sensu HM) under *T. javanica* once more.

Thus, while all lists recognize the name *T. alba*, they propose different treatments, diverging both in which concepts they recognize and in how they are ranked. Such complex cases are not uncommon in birds, and methods quantifying disagreement should be able to capture them as detailed as possible. To measure disagreement, Neate-Clegg *et al.* (2021) counted all instances of names that are not included in all four lists. Whenever disagreement about which names to recognize was not due to recent discoveries of unknown organisms (i.e., new discoveries not yet recognized by all authorities), they assumed it was due to disagreement about recognizing the group as a species or a subspecies of another group. In those cases, they lumped the disputed splits (i.e., both parent species and potential subspecies) into a single group about which there is disagreement. In the case of *Tyto alba*, where the four lists recognize respectively 1, 2, 3 and 4 species (not counting *T. glaucops*), Neate-Clegg *et al.* lump all these various treatments into a single name (*Tyto alba*) about which there is disagreement. Thus, due to the use of names (although perhaps not necessarily), Neate-Clegg *et al.* measure disagreement about *complexes of groups of birds* rather than about the delimitation of single groups of birds.

McClure *et al.* (2020), on the other hand, use a measurement method based on taxon concepts (instead of names). As discussed above, taxon concepts are groups of organisms as they are circumscribed and recognized by a particular source. As such, every grouping somehow recognized as a taxon within the *Tyto alba* complex represents a relevant taxon concept, whether specific authorities include them in their list or not. Such taxon concepts differ from taxon names in that what they refer to is unambiguously fixed, i.e., a particular circumscription of a group of organisms, typically given some unique identifier, regardless of rank and scientific name on this or that list. The main advantage of using concepts to study disagreement is that it makes it possible to precisely map the groups recognized by different authorities and evaluate whether there is disagreement. In the case of *Tyto alba*, referring just to the names tells you that there is disagreement about whether the group should be split, but it does not reveal whether the lists that split agree on how the split should be done. Comparing the concepts instead of the names reveals that, instead of there being a simple conflict between lumping and splitting, all four lists actually disagree about which groups of birds should be recognized as species and subspecies, with the status of 10 taxon concepts at stake. If one uses concepts to measure disagreement, each such instance of disagreement can be recorded and counted separately instead of being lumped into a single disputed split.

More precisely, there are three ways in which the use of concepts makes for a more precise measure of conflict than the use of names. First, the concepts methods captures the extent to which there is disagreement in in greater detail than the names method. This is because the names method cannot track how many distinct instances of disagreements there are, and consequently underestimates conflict. The fact that there are four different treatments for *T. alba* suggests that there are various, independent issues that experts disagree about. Lumping these into a single case of disagreement is not an accurate representation of how often the authorities agree or disagree. In addition, the names method cannot capture various degrees of disagreement between the four lists. Collapsing all disagreements about *T. alba* into a single disputed split makes it impossible to capture whether there is disagreement between one list and the other three, between each pair of lists separately, or any combination in between. The concepts method, on the other hand, counts instances of disagreements between list pairs separately. This way, it reveals that even though all lists include '*T. alba',* they all have different treatments and add

up to 6 cases of disagreement (i.e., one between each pair of lists). In many other cases, three lists agree and only one disagrees, signaling far weaker disagreement.

Second, the names approach cannot capture the associations of disagreement with biological variables (e.g., habitat, migration behavior) in the same detail. Because the names approach lumps multiple taxon concepts in one name, it must assume that the values for these variables are similar for all concepts subsumed under the name. But if it is a hypothesis that biological differences between concepts may well be the cause of conflict, lumping concepts and their biological properties comes with the risk of burying these causes or conflating the causes of substantially different cases. The concept approach, on the other hand, can attribute different characteristics to different concepts within a name or complex, and therefore distinguish between the various factors that are associated with these disagreements.

Neate-Clegg *et al.* (2021) raise the reverse worry that using concepts rather than names duplicates conflict and the variables that might explain it. However, disagreements such as those about the various groups in *T. alba* are real, and presumably each of the four lists has reasons why they recognize different groups of birds as species. Variables associated with each of these groups are not unduly duplicated as the different groups presumably have some different properties (why else would some lists split them?), and even if they do not, they still constitute distinct cases of disagreement. Indeed, counting all these different conflicts as one underestimates the extent to which bird specialists disagree about this group of birds and consequently also underestimates how often different causes of disagreement played a role.

Finally, the names method cannot capture that there are different kinds of conflict. Species classification consists of three importantly different steps: delimiting a concept, assigning a rank to the concept, and giving a name to the ranked concept (Fig. 1). Disagreement may arise at each of these steps, leading to conflicts that are different in kind and may have different causes and consequences. Using concepts is preferable over using names as it makes it possible to identify which kind of conflict there is about each group of birds. For example, in the case of *T. deroepstorffi*, there is a pure conflict of rank: all lists recognize the concept, but BL as subspecies and the other lists as species. On the other hand, regarding *T. glaucops*, there is a conflict of concepts: all lists recognize *T. glaucops* at species level, but BL includes populations under it, that other lists place under *T. alba* (or *T. javanica* in the case of IOC) (see Fig. 2). While the method advanced by Neate-Clegg *et al.* intentionally or unintentionally suggests that all conflicts are simply about whether groups ought to be recognized as species or as subspecies, the concepts method makes it possible to distinguish between conflict about rank, about delimitation, and about nomenclature. This is important as different kinds of conflict might be caused by different processes and in turn require different kinds of solutions.

In our study, we first distinguish between two broad kinds of disagreement: classificatory disagreement and nomenclatural disagreement. Classificatory disagreement is disagreement about how organisms should be organized into groups, and can be divided into disagreement about concepts and disagreement about ranking. Concept disagreements occur when one list recognizes a concept that does not occur on another list. Note that sometimes these different concepts go by the same name on different lists. In such cases, conflict is only visible if we use concepts rather than names. Rank disagreements, on the other hand, occur when two lists include the same taxon but give it a different rank, such as in the case of *T. deroepstorffi*.

Nomenclatural disagreement, then, is different from both rank and concept disagreement in that it is about the name given to groups of organisms. Nomenclatural disagreement occurs when two lists agree on the rank and delimitation of a concept but recognize the concept under a different name. This may be because of emendations, spelling differences, or different epithets. A special case of nomenclatural disagreement – when focusing on the species level – is where taxonomists agree on the name and rank

of a concept but place it in different genera. While this is strictly speaking also disagreement about classification (namely generic classification), it appears as a nomenclatural issue on species lists and, for most practical uses, the main impact of disagreement about generic classification lies in the fact that it changes the Linnaean names of these groups. Figure 3 summarizes these different kinds of conflict.

Thus, there are clear benefits to measuring conflict using concepts rather than names, and we see no good reason to rely on Linnaean names when a mapping of names into taxon concepts is available. Unfortunately, it is a lot of work to distill an exhaustive classification and mapping of taxon concepts from the taxonomic literature (which revolves around names), and there are very few groups for which an exhaustive database of taxon concepts exists. Fortunately, however, an exhaustive and open database of bird concepts does exist in the shape of Avibase (Lepage *et al.* 2014). In addition to including all bird concepts, nomenclatural and distribution data, Avibase indicates in which checklist each concept is included. It therefore has all the tools to measure disagreement about bird taxonomy. We argue that we should use this wealth of information and measure disagreement about bird taxonomy using concepts rather than names.

Methods

Measuring disagreement

We quantified disagreement in bird taxonomy by comparing the four most important and frequently used global lists of bird species at the time of starting this study: HM (4th edition, incl. corrigenda vol. 1–2), IOC (ver. 11.2), BL (ver. 5, Dec 2020) and CLEM (2019). The two previous studies on conflict suggest that a substantial portion of discrepancies between these lists may be caused by the fact that HM has not been updated since 2014 and does not take into account scientific research published after that date (McClure *et al.* 2020; Neate-Clegg *et al.* 2021). Because primary analysis showed that HM is not the main source of conflict among the four lists (see below) and because HM is still commonly used by important stakeholders such as natural history museums, we included it in the analysis, even though it has not been updated recently.

On February 25, 2021, we recorded from Avibase (https://avibase.bsc-eoc.org) all concepts listed as species on at least one of these four lists. For all of these concepts, we also recorded whether they are recognized as subspecies on one of the four lists. We then removed all concepts qualified as extinct from the database as the lists do not claim to cover all extinct bird species, and differences in the extinct concepts included by different lists thus may not necessarily indicate genuine taxonomic disagreement.
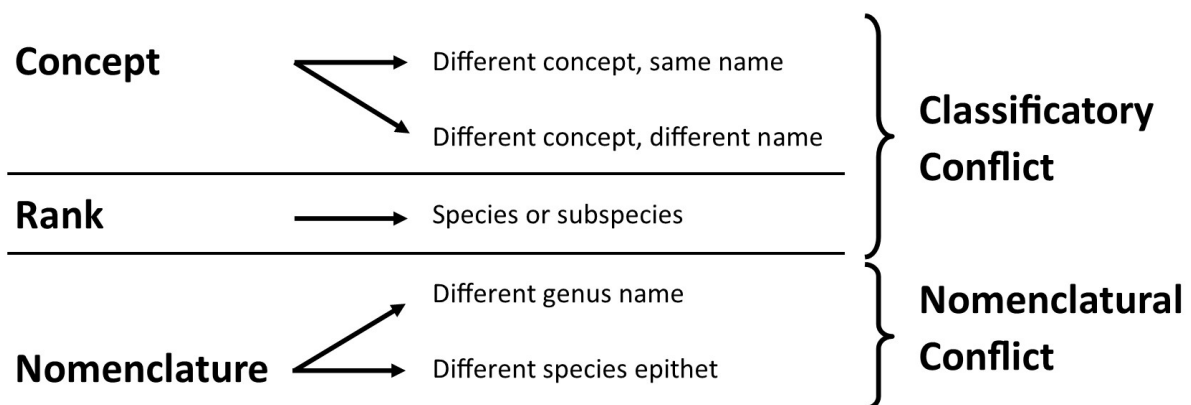

Fig. 3. The different kinds of conflict measured in this study. We measure rank conflict, concept conflict, nomenclatural conflict. Classificatory conflict is the combination of rank conflict and concept conflict.

As species and subspecies are the most important units in taxonomy, we did not include higher level taxa in the analysis.

For each concept recognized as a species on at least one list, we recorded for each pair of lists what kind of relation they have about that concept. This relation is either *agreement*, in case the concept is absent on both lists or present on both lists with the same rank, or one of the different kinds of conflict described above: *concept conflict*, *rank conflict*, or *nomenclatural conflict*. Concept conflict and rank conflict are mutually exclusive with regard to individual list-relations: rank conflict represents disagreement about the rank given to the exact same concept, so that both lists must recognize the very same concept before they can give it a different rank and hence be subject to rank conflict. The other forms of conflict are not mutually exclusive within list-relations (all rank conflicts entail nomenclatural conflict, and some cases of concept conflict also entail nomenclatural conflict). In those cases, we prioritized concept and rank conflict over nomenclatural conflict, and identified that particular list-relation as a form of classificatory conflict. This means that what we qualified as nomenclatural conflict, is pure nomenclatural conflict: there is agreement on the concept and on its rank, but not on its name.

Following this logic, we recorded six relations per conflict. From these relations, we derived four measures of conflict for each concept. First, a binary measure of classificatory conflict for which a concept scores '1' if there is concept-conflict or rank-conflict about that concept for at least one list-pair. Second, a binary measure of nomenclatural conflict for which a concept scores '1' in case there is nomenclatural conflict about that concept for at least one list-pair. Finally, ordinal measures of concept-conflict and rank-conflict, which for each concept track how many instances of rank and concept conflict there are about this concept. Because there are six list-pair relations per concept, the scores for the ordinal measures varied between 0, 3, 4, 5 and 6 (1 and 2 are logically impossible). As these are ordinal measures, we transformed these five possible values to 0–4 for the analysis.

Measuring research effort

To investigate the relation between research effort and taxonomic disagreement, we chose three metrics as proxies for research effort: the number of web pages on which a name occurs, measured by the number of Google search results; the number of scientific papers in which a name occurs, measured by Web Of Science search results; and the number of occurrences recorded for a name in GBIF. Unfortunately, these measures inevitably use names rather than concepts – taxa are mentioned on webpages and in scientific literature under names, not under concepts. Because in some cases names can be linked to more than one concept across lists, and in other cases more than one name can be linked to a concept, we limited the dataset for this part of the study to all names on the BL list. This limits the dataset (from 12 730 concepts to 10 999 concepts), but it ensures that name-based data can be unequivocally linked to concepts.

We assume the number of Google search results to be associated with research effort because Google search results for the scientific name of a species have been shown to be a good measure for the societal interest in the species (Correia *et al.* 2016, 2017; Ladle *et al.* 2019). Given that both research funding and the interests of individual taxonomists are influenced by societal salience and vice versa, more Google search results for a name should be associated with more research effort into that taxon. Indeed, previous research has demonstrated that scientific bird species names with more search results were typically described earlier, had extensive geographic ranges that intersect with technologically advanced societies, had more direct interactions with humans (such as hunting), and were more phenotypically conspicuous (Ladle *et al.* 2019). For all these factors, it seems likely that they are also associated with higher levels of scientific effort. In addition, at least for charismatic species (such as birds), there seems to be extensive overlap between societal and scientific interest (Jarić *et al.* 2019).

295

To get an estimate of the total number of search results for each name on the BL list, we used the Google API Client Library for Python (Google 2023) to use the Google API to do a global search for each name on 5 different occasions. The use of the API (Application Programming Interface) increases the chance that the search algorithm was not personalized to search history and location, as would be the case if searches were carried out through a web browser (Hannak *et al.* 2013). However, because the exact algorithm for estimating the number of results is a commercial secret, it is unclear to what extent the searches were effectively fully de-personalized (Uyar 2009; Ladle *et al.* 2019). We used quoted search strings (e.g., "*Tyto alba*") to restrict the results to exact matches. Because many bird species are known under many different synonyms, and the inclusion of such synonyms has been shown to strongly affect the number of Google search results for bird species (Correia *et al.* 2018), we retrieved all synonyms for each birdlife species from its Avibase page, and included all web pages on which at least one of the synonyms occurred. Because there were small differences between different runs in the estimated number of results for many of these searches – in particular for names with a low number of results (see Supp. file 1: Fig. S1) – we repeated the search 5 times and used the mean number of search results for the statistical models.

Clarivate's Web of Science (WoS; https://www.webofknowledge.com/) is one of the largest databases of scientific publications, citations and other bibliometric data. Among other things, it indexes over 21 000 high quality scholarly journals and their publications across all fields of academia. We assume the number of WoS search results to be associated with research effort because each hit in WoS corresponds to a scientific publication using that name. Presumably, more such hits indicate that more research into that name or close relatives has been done, and that more data on the taxon is available. Note that the search in WoS is expected to cover both taxonomic effort through specialized taxonomic journals and research effort in other fields such as ecology or evolutionary biology. We used the WOS package (Bacis 2021) in python to perform searches through the WoS search API for each of the quoted species names and their synonyms in the abstracts, titles and keywords of all publications in the WoS Core Collection, and recorded the number of hits for each of these.

The Global Biodiversity Information Facility (GBIF; https://www.gbif.org/) is a platform that makes scientific data on biodiversity openly, freely and easily available. Among other data, it contains large occurrence datasets linked to scientific species names. We assume the number of such occurrences recorded in GBIF for a name to be associated with the amount of research effort into the taxon with that name. The relation here is more straightforward than for the WoS and Google search results as GBIF occurrences are data points and hence more occurrences directly indicate that more data is available. We used the GBIF occurrence API (https://www.gbif.org/developer/occurrence) to obtain the number of occurrences for each of the species names on the BL list. Unlike the Google and WoS searches, we did not include the Avibase synonyms in these searches as GBIF has its own database of synonyms and allows searches to include known synonyms.

It should be noted here that none of these three measures of research effort is perfect. The algorithm through which Google estimates the total number of hits is not public, and the number of search results for each name varied between different iterations of the search even though the procedure was not changed (see Supp. file 1: Fig. S1). Moreover, some web pages might not be included in the results as the internet coverage of Google's index is incomplete (Uyar 2009). Hence, it is difficult to evaluate how reliable the estimates are. Similarly, WoS search results are an imperfect measure of research effort because the coverage of WoS is far from perfect. In particular, WoS is biased towards more recent, English speaking journals with higher impact factors. As there are many small journals, monographs, and older papers in taxonomy that are not indexed by WoS, searches in the WoS database are likely to miss some of the research effort. Finally, GBIF occurrences are an imperfect measure of effort because not all occurrences are associated with the same kind and degree of research effort. In particular for birds,

much of the occurrence data is from platforms such as eBird (Auer *et al.* n.d.), which enable amateurs to register when they have observed a species. Such citizen science data may well be of lower quality than data collected by professionals, and is biased towards certain regions and habitats (Callaghan *et al.* 2021; Johnston *et al.* 2023; Scher & Clark 2023).

It should be remembered that all three measures inevitably relied on the use of names rather than concepts. Hence, they are subject to some of the same limitations of the names approach that we discussed above. Most importantly, names are often ambiguous, and it is far from clear how many of the search results or occurrences are actually connected to the particular taxon concept on the BL list rather than some other taxon that sometimes goes by the same name.

Statistical analysis

The directed acyclic graph (DAG) in Fig. 4 summarizes what we assume to be the causal relations between the three proxies for research effort and research effort itself. In addition, the DAG also shows the assumed causal role of species concepts, and (recent) diversification and other biological properties of the taxa, in generating taxonomic disagreement. Using the so-called "backdoor criterion" (Cinelli *et al.* 2022), we used this DAG to select the variables to include in the regression analyses to estimate the causal influence of effort on disagreement. The backdoor criterion helps to identify a set of variables in a DAG that, when controlled for, allow you to estimate the causal effect from observational data by blocking all backdoor paths between the treatment and outcome variables that could bias the estimate. Note that a different DAG might require different variables to be included in the model, and consequently could lead to different estimates. Assuming this DAG, then, we ran six Bayesian ordered logistic regressions (Kruschke 2014; McElreath 2016) with respectively Google search results, WoS search results, and GBIF occurrences as the predictor variable and the ordinal measures of concept conflict and rank conflict as the outcome variables. Because we assumed WoS searches without results and GBIF searches without occurrences to be qualitatively different from one or more search results, we modeled the zeros for these models through a binary variable separately from the number of occurrences or hits. All non-zero GBIF and WoS results and all Google search results were transformed using the base-10 logarithm as the data was heavily skewed. As there were no empty Google search results, the regressions with Google hits as the predictor variable did not include a separate binary variable for empty searches. We used weakly informative priors in all models, based on the results of the pilot study. All the analyses were accomplished using Markov chain Monte Carlo methods (MCMC; (van Ravenzwaaij *et al.* 2018)). For the detailed specification of all the models, see Supp. file 4: S4 Files.
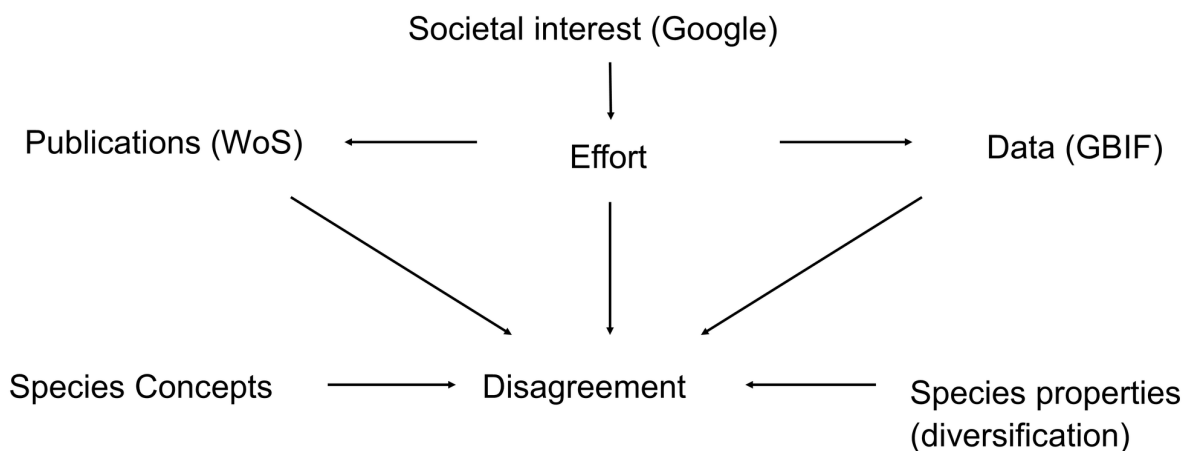

Fig. 4. Directed acyclic graph (DAG) for the effort analysis. This DAG shows the assumed causal relations between disagreement, the proxies for research effort, and the three factors typically assumed to influence disagreement (species concepts, effort and diversification).

All analyses were done in Jupyter notebooks in Python using Pandas (McKinney 2010) and Numpy (Harris *et al.* 2020), and Seaborn (Waskom *et al.* 2022), Matplotlib (Hunter 2007) and Arviz (Kumar *et al.* 2019) for plotting results. All statistical analysis was done using the Pymc (Salvatier *et al.* 2016) library in Python (see Supp. file 4: S4 Files for all package versions). All data are made available through Zenodo (Conix & Cuypers 2023).

Results

Distribution of conflict and conflict types

After removing extinct taxa, we recorded 12 730 unique taxon concepts recognized as species by at least one list, with a difference of almost 1000 species between the list that recognizes the lowest number of species (HM, 10 038 species) and the list that recognizes the highest number of species (BL, 10 999 species). The lists agreed fully about 7874 (61.85%) of these unique concepts.

Classificatory conflict clearly outnumbered nomenclatural conflict: there was classificatory conflict about 3974 concepts (31.22%), and nomenclatural conflict about 1023 concepts (8.04%) (Fig. 5a). Of the nomenclatural conflict, most (83.70%) was conflict about the genus a concept should be placed in. Thus, there was very little conflict concerning specific epithets. Of the classificatory conflict, 63.81% were conflicts about which concepts to recognise, while 38.35% were conflicts about rank (with a small proportion of concepts about which there was both rank and concept conflict). Thus, concept conflict (19.92% of all concepts) clearly outnumbered rank conflict (11.97% of all concepts). Classificatory conflict was lower (16.86%) when counted by list-relations than by concepts (Fig. 5b), and the ordinal conflict scores showed that '3 lists vs 1 list' (78.76%) is by far the most common configuration of classificatory conflict in these list relations, followed by '2 vs 2' (19.68%) and only a small proportion of '2 vs 1 vs 1' (2.16%) (1 vs 1 vs 1 vs 1 was not logically possible).

The proportions of the different kinds of conflict were similar across pairs of the four lists, and reflect the overall proportions reported above. Comparing different lists, disagreement was clearly highest between
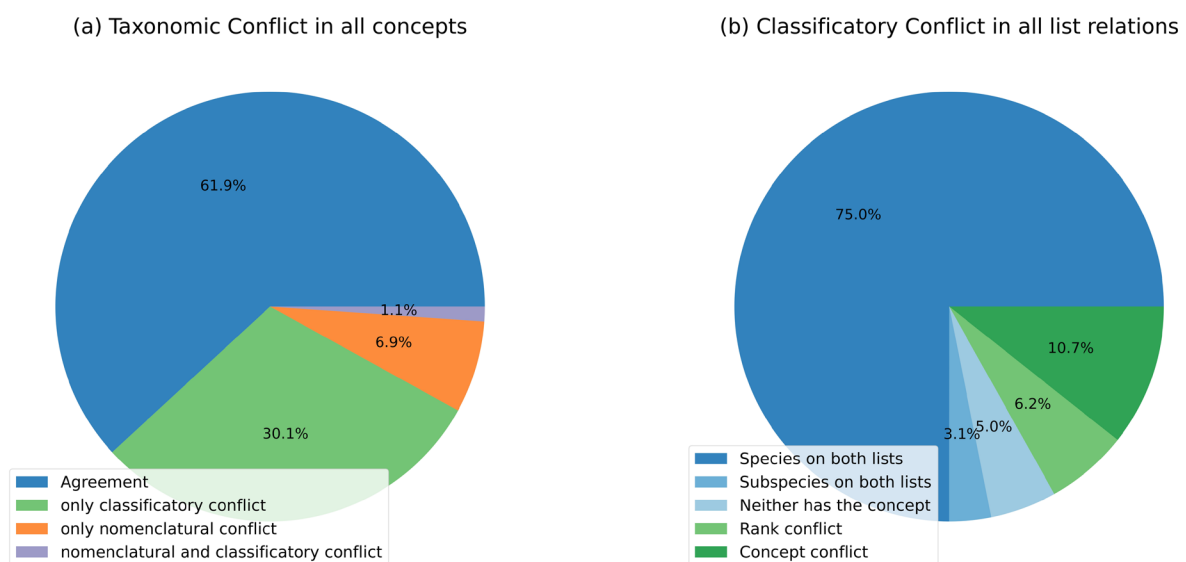

Fig. 5. Prevalence of conflict in bird lists. **a**. Prevalence of taxonomic conflict and agreement across the 12 730 concepts listed by at least one of the four lists. **b**. Prevalence of classificatory conflict and agreement across the 76 380 list-relations. Agreement relations are in shades of blue and conflict relations are in shades of green.

Table 1. Comparison of lists and how they contribute to conflict.

	Species concept	#species	#subspecies in dataset	#unique concepts	#unique species	#unique opinions	%tax agreement if removed
BL	BSC	10999	345	456	1037	1712	80.7%
HM	BSC	10038	1362	392	417	1370	77.9%
IOC	ESC	10824	655	182	319	661	72.0%
CLEM	BSC	10569	866	66	81	366	69.7%

BL and HM (24.12%) and lowest between IOC and CLEM (7.50%). BL and HM were shown to be the greatest contributors to classificatory conflict: classificatory agreement between all lists goes up from 68.78% to 80.50% if BL is removed, and to 77.54% if HM is removed (Table 1). This was also reflected in the number of unique concepts, unique species and unique classificatory opinions that each list has (Table 1). CLEM scored lowest for each of these, with high scores for BL and HM. In particular, BL had many species (1037) that none of the other lists recognize as species (e.g., 81 for CLEM), and HM had many subspecies (1362) compared to other lists (e.g., 345 for BL).

Research effort

For all three measures of research effort, an increase in research effort was associated with decreased levels of rank conflict. That is, a decrease in Google search results (-1.77, 97% HDI = [-1.87, -1.67]), WoS search results (-0.38, 97% HDI = [-0.50, -0.25]) or GBIF occurrences (-0.69, 97% HDI = [-0.74, -0.64]) was associated with a higher level of rank disagreement, as was an increase in names without WoS search results (0.63, 97% HDI = [0.47, 0.77]) or GBIF occurrences (0.23, 97% HDI = [-0.04, 0.50]; see Supp. file 2: Table S2 for all regression coefficients). Concept conflict, on the other hand, was stable or increased slightly with increased research effort for Google search results (0.17, 97% HDI = [0.08,
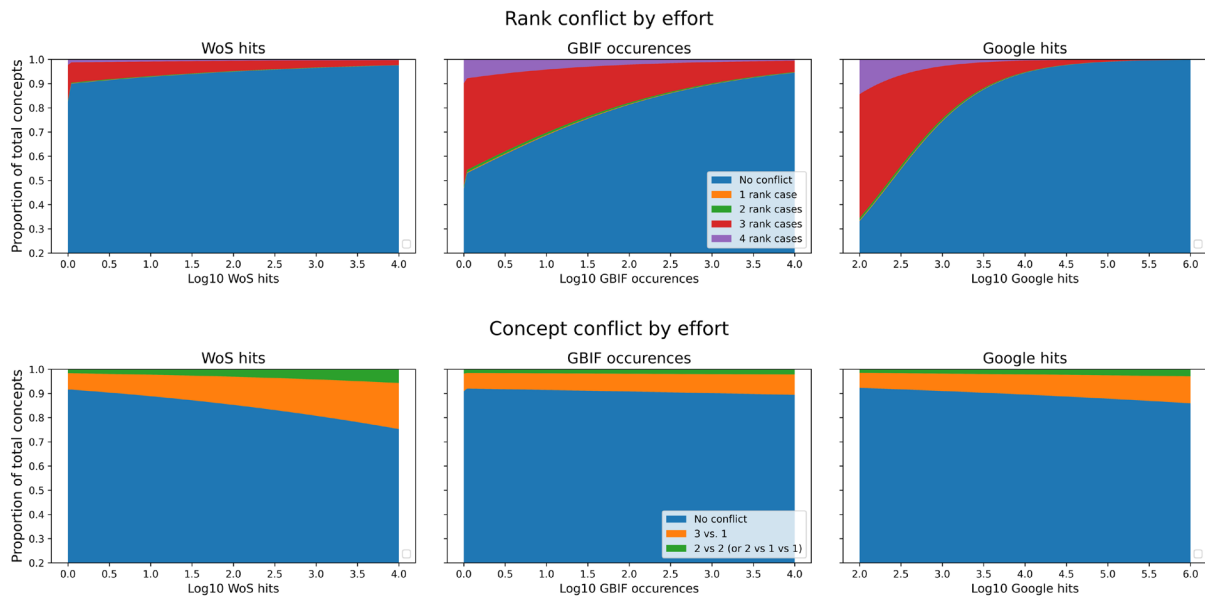

Fig. 6. Rank and concept conflict by the three measures of research effort. These plots show the posterior predictive proportions for each of the ordinal scores of rank conflict (top) and concept conflict (bottom) as research effort increases. Note that the x-axis has a different range for the Google hits plots (right column) than for the other two, as the realistic values for Google hits have a substantially different range.

0.26]), WoS search results (0.32, 97% HDI = [0.24, 0.41]) and GBIF occurrences (0.08, 97% HDI = [0.0.3, 0.13]; see Supp. file 2: Table S2 for all coefficients). Figure 6 shows how the posterior predictive proportions of each of the ordinal scores of rank and concept conflict change as effort (as tracked by each of the measures) increases. Note that for rank conflict, 3 cases of rank conflict occurred far more often than 1, 2 or 4 cases of rank conflict. These are all cases where three lists agreed on the status of a concept (either as a species or subspecies), and one list included the taxon with the other rank. Taxon concepts with 1, 2 and 4 cases of rank conflict were all concepts about which there was less agreement (2 vs 2 or 2 vs 1 vs 1 rather than 3 vs 1), which occurred far less often.

Discussion

Patterns of disagreement in bird classification

Our results confirm that avian classification is subject to substantial disagreement. Even if nomenclatural conflict is ignored, there is disagreement about more than one in four concepts that are recognized as species by one of the four global checklists. This rate of conflict remains high (22%) even if the relatively outdated HM list is removed from the analysis. This high rate of conflict is somewhat mitigated by a lower rate of conflict among list relations, i.e., by the fact that in most cases of classificatory conflict there were 3 lists with the same opinion opposing one dissident list. Still, our results clearly confirm that disagreement in bird taxonomy is a substantial and potentially far-reaching problem that is highly likely to affect policy-makers and other users of these lists. As such, our results add further urgency to calls to resolve the problem of taxonomic disorder (Agapow *et al.* 2004; Isaac *et al.* 2004; Garnett *et al.* 2020).

It is often assumed that taxonomic disagreement boils down to a simple opposition between lumping and splitting (Agapow *et al.* 2004; Isaac *et al.* 2004; Neate-Clegg *et al.* 2021), but our data show that this is not the case. In addition to such ranking disagreement – and more prevalent than it – there is disagreement about which organisms should be grouped into recognized taxa to begin with. The high prevalence of such conflict about which groups to recognize (regardless of their rank) is worrying because it is more likely than rank conflict to have practical implications. For example, in conservation policies, discussions about ranks can be bypassed by allowing the listing of subspecies and other non-species entities, as is the case in the United States (Wheeler 2014). However, if there is disagreement about the boundaries of relevant taxa, such a strategy is not possible as it is unclear which organisms a name refers to. This forces policy-makers and conservationists to make the difficult taxonomic decisions of which concepts to work with when taxonomists disagree.

Of course, not all cases of disagreement between lists are likely to reflect fundamental differences in taxonomic opinions. Some conflict, and perhaps even a substantial part of all conflict, is undoubtedly due to delays in incorporating new evidence, in simple oversight, or because different lists use different sets of evidence. For example, a recent study shows that many of the unique BL species accepted on the basis of the Tobias-criteria are confirmed by follow-up research (Tobias *et al.* 2021). Part of the discrepancies between BL and the other lists may thus lie in the fact that these lists have not considered the same evidence, or only accept new species based on peer reviewed publications. In cases of ongoing speciation, it may also be that different lists assign different ranks to a taxon but fundamentally agree that the ranking decision is arbitrary. In that case, (rank) conflict is an artefact of taxonomists having to apply a binary classification system to variation that is often continuous. However, we believe that even all these seemingly superficial cases of conflict are important, as they lead to differences between lists and thus have practical consequences. For example, much conservation legislation worldwide does only grant protection to species-level taxa (Garnett & Christidis 2007). Whenever that is the case, apparently arbitrary ranking decisions of these lists could have life-or-death consequences for those taxa.

Disagreement and research effort

The models associating proxies of research effort to rank and concept conflict show that more research effort is associated with lower levels of rank conflict and stable or slightly higher levels of concept conflict. This suggests that the relation between taxonomic disagreement and research effort is more complex than the simple hypothesis proposed by Neate-Clegg *et al.* and McClure *et al.*, which proposes that disagreement about taxa diminishes as we come to know more about these taxa. Instead, one kind of conflict seems to decrease with research effort, while another increases or is unaffected by it.

To understand why concept conflict may be unaffected by increased research effort – and to understand why it is so prevalent in the first place – it is important to understand what typical cases of concept conflict about birds look like. To check this, we randomly selected 10 cases of concept conflict from the dataset (see Supp. file 3: Table S3). Detailed investigation of these cases suggests that concept conflict typically concerns complexes in which there is broad agreement about which basic units or (meta) populations to recognize. These basic units are often below the subspecies level. Concept conflict, then, may often be the disagreement about how to group these basic units into species and subspecies. This is also true for the case of *Tyto alba* as discussed above: generally accepted (meta)populations are grouped in diverging ways in subspecies and species, resulting in many different concepts across treatments by the different lists (Fig. 2).

This pattern – consensus about the populations but not about how these should be grouped into species and subspecies – may explain why concept conflict does not decrease, and may even increase, with research effort. Substantive understanding of a taxon on the population-level requires extensive study of the group, and is probably only available for groups about which a lot of data has been gathered. While such extensive understanding of a taxon undoubtedly clears up potential mistakes or misunderstandings that were due to a lack of information, it may also uncover new complexities or ways in which various kinds of data and grouping criteria are in conflict. Indeed, taxonomic controversies are often caused by the fact that different types of data contradict each other (De Queiroz 2005), and there are plenty of examples where more research and data increase the chances of conflicting evidence and analyses (Satler *et al.* 2013). This is particularly the case when deciding on the boundaries of concepts. If only a few specimens of an understudied group are available, they might show clear differences and be easily classified in taxon concepts, the main question then being which rank to give them. However, if more and more specimens become available, these may show more subtle variations. This increased complexity of the patterns in the studied characters might then move the discussion from the – relatively operational – question of ranks to the question of which groups form relevant taxa, and which do not.

Note that this potential explanation is largely speculative, and more fine-grained research into the focus of taxonomists on both concepts and ranks throughout research processes is needed. However, our results do suggest clearly that it is wrong to assume a simple inverse relation between research effort and taxonomic disagreement. Neate-Clegg *et al.* (2021) raise the suggestion that taxonomic disagreement is a transient phenomenon, caused in particular by the existence of grey area species from groups with high diversification rates. On this hypothesis, taxonomic disagreement can be resolved simply by increasing research effort. The results of our analysis paint a more complicated picture, on which research effort can both decrease and increase taxonomic conflict. Indeed, there may be a 'sweet spot' of research effort that clears up disagreements due to a severe lack of information, but that does not yet uncover all the evolutionary complexities that play on the level of populations. At that intermediate level of knowledge, there is more likely to be agreement on the specific or subspecific level of groups rather than the population-level, and there less likely to be enough information to have disagreements about which populations should be grouped into which species or subspecies.

Of course, we do not wish to say that more information about organisms is not desirable in itself. Rather, the point is that resolving taxonomic disagreement may be more complex than simply increasing research effort, and other solutions – such as taxonomic governance or the development and use of taxon concept-based databases – may be needed as well. In particular, we think that research effort should be distributed in a coordinated and evidence-based manner, in particular given limitations in the resources available for taxonomy. For example, if our 'sweet spot hypothesis' above were true, taxonomists should focus on those taxa that have not yet reached that sweet spot, and be careful not to spend ever more resources on well-studied taxa on which there already is conflict.

Conclusion and prospects

We want to highlight three main take-aways of this study. First, our results confirm the value of using concepts rather than names, both for taxonomic work and for measuring taxonomic disagreement. More precisely, our results suggests that the use of concepts for measuring conflict can pick up on many cases of conflict that remain unnoticed if conflict is measured using names. In addition, the use of concept makes it possible to distinguish between rank conflict and concept conflict. These two kinds of disagreement are different problems subject to different drivers, and hence may well require different solution. Thus, publishers of taxonomic data should be encouraged to publish their data using concepts, and databases should – where possible – try to follow the example of Avibase and structure their data around concepts rather than names.

Second, our results point to the efficient distribution of research effort as an important focus for future research. Previous claims about the relation between research effort and taxonomic disagreement were probably too quick, and further research is needed to find out how exactly research effort affects disagreement, and how disagreements evolve along with research processes through time. A better picture of how research effort influences taxonomic disagreement can then be used to choose which taxa to spend more taxonomic effort on.

Finally, our study shows that even in one of the best-studied groups (or, perhaps, *because* birds are one of the best studied groups), disagreement and confusion about the classification of organisms is rampant. Because these classifications play a crucial role in conservation, this is likely to have far-reaching consequences in practice. We therefore applaud initiatives like the Working Group on Avian Taxonomy to create a consensus bird list and urge list-makers and taxonomists in other charismatic taxa to work towards such a consensus list as well. This does not mean that there is no place for taxonomic disagreement (Sterner *et al.* 2020). We believe that such disagreements are inevitable and are often a motor for scientific progress. However, these disagreements should be coordinated, tractable and productive. For example, they can be incorporated in a pragmatic consensus-list of species, in the shape of statements about uncertainty, alternative classifications, and the identification of taxa that require more research.

Authorship contributions

Stijn Conix: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing.

Vincent Cuypers: Conceptualization, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing.

Charles H. Pence: Conceptualization, Writing – Review & Editing.

Acknowledgements

Disclosure statement

The authors declare that there is no conflict of interest.

References

Agapow P.-M., Bininda-Emonds O.R.P., Crandall K.A., Gittleman J.L., Mace G.M., Marshall J.C. & Purvis A. 2004. The impact of species concept on biodiversity studies. *The Quarterly Review of Biology* 79 (2): 161–179. https://doi.org/10.1086/383542

Auer T., Barker S., Borgmann K., Charnoky M., Childs D., Curtis J., Davies I., Downie I., Fink D., Fredericks T., Ganger J., Gerbracht J., Hanks C., Hochachka W., Iliff M., Imani J., Johnston A., Lenz T., Levatich T., ... & Wood C. undated. *EOD – eBird Observation Dataset*. https://doi.org/10.15468/aomfnb

Bacis E. 2021. WOS. Ver. 0.2.7. Available from https://pypi.org/project/wos/ [accessed 16 May 2024].

Berendsohn W.G. 1995. The concept of "potential taxa" in databases. *Taxon* 44 (2): 207–212. https://doi.org/10.2307/1222443

Birdlife International 2020. Handbook of the Birds of the World and BirdLife International digital checklist of the birds of the world. Version 5. Available from http://datazone.birdlife.org/userfiles/file/Species/Taxonomy/HBW-BirdLife_Checklist_v5_Dec20.zip [accessed 16 May 2024].

Callaghan C.T., Poore A.G.B., Hofmann M., Roberts C.J. & Pereira H.M. 2021. Large-bodied birds are over-represented in unstructured citizen science data. *Scientific Reports* 11 (1): 19073. https://doi.org/10.1038/s41598-021-98584-7

Cinelli C., Forney A. & Pearl J. 2022. A crash course in good and bad controls. *Sociological Methods & Research*: 00491241221099552. https://doi.org/10.1177/00491241221099552

Clements J.F., Schulenberg T.S., Iliff M.J., Billerman, S.M., Fredericks T.A., Sullivan B.L. & Wood C.L. 2019. The eBird/Clements Checklist of Birds of the World: v2019. Available from https://www.birds.cornell.edu/clementschecklist/introduction/updateindex/overview-august-2019/2019-citation-checklist-download/ [accessed 16 May 2024].

Conix S. & Cuypers V. 2023. Dataset for 'Measuring and explaining disagreement about bird taxonomy'. https://doi.org/10.5281/zenodo.10078307

Conix S., Garnett S.T., Thiele K.R., Christidis L., van Dijk P.P., Bánki O.S., Barik S.K., Buckeridge J.S., Costello M.J., Hobern D., Kirk P.M., Lien A., Nikolaeva S., Pyle R.L., Thomson S.A., Zhang Z.-Q. & Zachos F.E. 2021. Towards a global list of accepted species III. Independence and stakeholder inclusion. *Organisms Diversity & Evolution* 21: 631–643. https://doi.org/10.1007/s13127-021-00496-x

Correia R.A., Jepson P.R., Malhado A.C.M. & Ladle R.J. 2016. Familiarity breeds content: assessing bird species popularity with culturomics. *PeerJ* 4: e1728. https://doi.org/10.7717/peerj.1728

Correia R.A., Jepson P., Malhado A.C.M. & Ladle R.J. 2017. Internet scientific name frequency as an indicator of cultural salience of biodiversity. *Ecological Indicators* 78: 549–555. https://doi.org/10.1016/j.ecolind.2017.03.052

Correia R.A., Jarić I., Jepson P., Malhado A.C.M., Alves J.A. & Ladle R.J. 2018. Nomenclature instability in species culturomic assessments: why synonyms matter. *Ecological Indicators* 90: 74–78. https://doi.org/10.1016/j.ecolind.2018.02.059

Cuypers V., Reydon T.A.C. & Artois T. 2022. Deceiving insects, deceiving taxonomists? Making theoretical sense of taxonomic disagreement in the European orchid genus *Ophrys*. *Perspectives in Plant Ecology, Evolution and Systematics* 56: 125686. https://doi.org/10.1016/j.ppees.2022.125686

De Queiroz K. 2005. A unified concept of species and its consequences for the future of taxonomy. *Proceedings of the California Academy of Sciences* 56 (supp. 1): 196–215.
Available from https://repository.si.edu/handle/10088/4507 [accessed 14 May 2024].

Dickinson E. & Christidis L. 2014. *The Howard and Moore Complete Checklist of the Birds of the World.* Fourth Edition. Aves Press, Eastbourne, UK.

Franz N., Peet R.K. & Weakley A. 2008. On the use of taxonomic concepts in support of biodiversity research and taxonomy. *In*: Wheeler Q.D. (ed.) *The New Taxonomy*: 63–87. Systematics Association Special Volume, CRC Press, Boca Raton. https://doi.org/10.1201/9781420008562

Franz N.M., Pier N.M., Reeder D.M., Chen M., Yu S., Kianmajd P., Bowers S. & Ludäscher B. 2016. Two influential primate classifications logically aligned. *Systematic Biology* 65 (4): 561–582. https://doi.org/10.1093/sysbio/syw023

Garnett S.T. & Christidis L. 2007. Implications of changing species definitions for conservation purposes. *Bird Conservation International* 17 (3): 187–195. https://doi.org/10.1017/S0959270907000809

Garnett S.T. & Christidis L. 2017. Taxonomy anarchy hampers conservation. *Nature* 546 (7656): 25–27. https://doi.org/10.1038/546025a

Garnett S.T., Christidis L., Conix S., Costello M.J., Zachos F.E., Bánki O.S., Bao Y., Barik S.K., Buckeridge J.S., Hobern D., Lien A., Montgomery N., Nikolaeva S., Pyle R.L., Thomson S.A., Dijk P.P. van, Whalen A., Zhang Z.-Q. & Thiele K.R. 2020. Principles for creating a single authoritative list of the world's species. *PLOS Biology* 18 (7): e3000736. https://doi.org/10.1371/journal.pbio.3000736

Gill F., Donsker D. & Rasmussen P.C. (eds) 2021. *IOC World Bird List (v11.2)*.
https://doi.org/10.14344/IOC.ML.11.2

Google 2023. Google API Client Library for Python.
Available from https://github.com/googleapis/google-api-python-client [accessed 16 May 2024].

Hannak A., Sapiezynski P., Molavi Kakhki A., Krishnamurthy B., Lazer D., Mislove A. & Wilson C. 2013. Measuring personalization of web search. *In*: *Proceedings of the 22nd International Conference on World Wide Web*: 527–538. Association for Computing Machinery, New York, NY, USA.

Harris C.R., Millman K.J., van der Walt S.J., Gommers R., Virtanen P., Cournapeau D., Wieser E., Taylor J., Berg S., Smith N.J., Kern R., Picus M., Hoyer S., van Kerkwijk M.H., Brett M., Haldane A., del Río J.F., Wiebe M., Peterson P., ... & Oliphant T.E. 2020. Array programming with NumPy. *Nature* 585 (7825): 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter J.D. 2007. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55

Isaac N.J.B., Mallet J. & Mace G.M. 2004. Taxonomic inflation: its influence on macroecology and conservation. *Trends in Ecology & Evolution* 19 (9): 464–469. https://doi.org/10.1016/j.tree.2004.06.004

Jarić I., Correia R.A., Roberts D.L., Gessner J., Meinard Y. & Courchamp F. 2019. On the overlap between scientific and societal taxonomic attentions — Insights for conservation. *Science of the Total Environment* 648: 772–778. https://doi.org/10.1016/j.scitotenv.2018.08.198

Johnston A., Matechou E. & Dennis E.B. 2023. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution* 14 (1): 103–116. https://doi.org/10.1111/2041-210X.13834

Kruschke J. 2014. *Doing Bayesian Data Analysis: a Tutorial with R, JAGS, and Stan*. Academic Press.

Kumar R., Carroll C., Hartikainen A. & Martin O. 2019. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software* 4 (33): 1143. https://doi.org/10.21105/joss.01143

Ladle R.J., Jepson P., Correia R.A. & Malhado A.C.M. 2019. A culturomics approach to quantifying the salience of species on the global internet. *People and Nature* 1 (4): 524–532. https://doi.org/10.1002/pan3.10053

Lepage D., Vaidya G. & Guralnick R. 2014. Avibase – a database system for managing and organizing taxonomic concepts. *ZooKeys* 420: 117–135. https://doi.org/10.3897/zookeys.420.7089

Lien A.M., Conix S., Zachos F.E., Christidis L., van Dijk P.P., Bánki O.S., Barik S.K., Buckeridge J.S., Costello M.J., Hobern D., Montgomery N., Nikolaeva S., Pyle R.L., Thiele K., Thomson S.A., Zhang Z.-Q. & Garnett S.T. 2021. Towards a global list of accepted species IV: overcoming fragmentation in the governance of taxonomic lists. *Organisms Diversity & Evolution* 21: 645–655. https://doi.org/10.1007/s13127-021-00499-8

McClure C.J.W., Lepage D., Dunn L., Anderson D.L., Schulwitz S.E., Camacho L., Robinson B.W., Christidis L., Schulenberg T.S., Iliff M.J., Rasmussen P.C. & Johnson J. 2020. Towards reconciliation of the four world bird lists: hotspots of disagreement in taxonomy of raptors. *Proceedings of the Royal Society B: Biological Sciences* 287 (1929): 20200683. https://doi.org/10.1098/rspb.2020.0683

McElreath R. 2016. *Statistical Rethinking: a Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, Boca Raton.

McKinney W. 2010. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*: 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Neate-Clegg M.H.C., Blount J.D. & Şekercioğlu Ç.H. 2021. Ecological and biogeographical predictors of taxonomic discord across the world's birds. *Global Ecology and Biogeography* 30 (6): 1258–1270. https://doi.org/10.1111/geb.13300

van Ravenzwaaij D., Cassey P. & Brown S.D. 2018. A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review* 25 (1): 143–154. https://doi.org/10.3758/s13423-016-1015-8

Salvatier J., Wiecki T.V. & Fonnesbeck C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* 2: e55. https://doi.org/10.7717/peerj-cs.55

Satler J.D., Carstens B.C. & Hedin M. 2013. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, Aliatypus). *Systematic Biology* 62 (6): 805–823. https://doi.org/10.1093/sysbio/syt041

Scher C.L. & Clark J.S. 2023. Species traits and observer behaviors that bias data assimilation and how to accommodate them. *Ecological Applications* 33 (3): e2815. https://doi.org/10.1002/eap.2815

Sterner B., Witteveen J. & Franz N. 2020. Coordinating dissent as an alternative to consensus classification: insights from systematics for bio-ontologies. *History and Philosophy of the Life Sciences* 42 (1): 8. https://doi.org/10.1007/s40656-020-0300-z

Thiele K.R., Conix S., Pyle R.L., Barik S.K., Christidis L., Costello M.J., van Dijk P.P., Kirk P., Lien A., Thomson S.A., Zachos F.E., Zhang Z.-Q. & Garnett S.T. 2021. Towards a global list of accepted species I. Why taxonomists sometimes disagree, and why this matters. *Organisms Diversity & Evolution* 21: 615–622. https://doi.org/10.1007/s13127-021-00495-y

Thomson S.A., Thiele K., Conix S., Christidis L., Costello M.J., Hobern D., Nikolaeva S., Pyle R.L., van Dijk P.P., Weaver H., Zachos F.E., Zhang Z.-Q. & Garnett S.T. 2021. Towards a global list of accepted species II. Consequences of inadequate taxonomic list governance. *Organisms Diversity & Evolution* 21 (4): 623–630. https://doi.org/10.1007/s13127-021-00518-8

Tobias J.A., Donald P.F., Martin R.W., Butchart S.H.M. & Collar N.J. 2021. Performance of a points-based scoring system for assessing species limits in birds. *Ornithology* 138 (2): ukab016. https://doi.org/10.1093/ornithology/ukab016

Trimble M.J. & van Aarde R.J. 2012. Geographical and taxonomic biases in research on biodiversity in human-modified landscapes. *Ecosphere* 3 (12): art119. https://doi.org/10.1890/ES12-00299.1

Uva V., Päckert M., Cibois A., Fumagalli L. & Roulin A. 2018. Comprehensive molecular phylogeny of barn owls and relatives (Family: Tytonidae), and their six major Pleistocene radiations. *Molecular Phylogenetics and Evolution* 125: 127–137. https://doi.org/10.1016/j.ympev.2018.03.013

Uyar A. 2009. Investigation of the accuracy of search engine hit counts. *Journal of Information Science* 35 (4): 469–480. https://doi.org/10.1177/0165551509103598

Waskom M., Gelbart M., Botvinnik O., Ostblom J., Hobson P., Lukauskas S., Gemperline D.C., Augspurger T., Halchenko Y., Warmenhoven J., Cole J.B., Hoeven E. ter, Ruiter J. de, Vanderplas J., Hoyer S., Pye C., Miles A., Swain C., Meyer K., ... & O'Kane D. 2022. *mwaskom/seaborn: v0.12.2 (December 2022)*. Zenodo. https://doi.org/10.5281/zenodo.7495530

Wheeler T. 2014. The concept of species with constant reference to killer whales. *Washington Journal of Environmental Law and Policy* 4 (1): 250–284.

Supplementary files

Supp. file 1. Fig. S1. Variation between multiple repeats of the same google searches. The log10 of the mean of the 5 google searchers for each species name by the coefficient of variation (CV) of the estimated number of search results for each of these searches.
https://doi.org/10.5852/ejt.2024.943.2601.11889

Supp. file 2. Table S2. Regression coefficients of all statistical models associating disagreement to proxies of taxonomic effort. https://doi.org/10.5852/ejt.2024.943.2601.11891

Supp. file 3. Table S3. List of cases of concept conflict manually checked. We selected these cases randomly using the Python library Numpy. Because concept conflict was unexpectedly high, we confirmed manually whether these were all cases concept conflict and investigated each case in detail. https://doi.org/10.5852/ejt.2024.943.2601.11893

Supp. file 4. Files S4. html exports of the jupyter notebooks used for the statistical analyses in this paper. The code can also be consulted on the github repository of the paper.
https://doi.org/10.5852/ejt.2024.943.2601.11899