OXFORD

BioScience, 2023, **73**, 728–737
https://doi.org/10.1093/biosci/biad081
Advance access publication date: 28 September 2023
Thinking of Biology

Taxonomic disagreement about ranks in gray-area taxa: A vignette study

Stijn Conix, Vincent Cuypers, Frank E. Zachos and Andreas De Block

Stijn Conix (stijn.conix@uclouvain.be) is a philosopher of science at the Université Catholique de Louvain-La-Neuve, in Louvain-La-Neuve, Belgium. Vincent Cuypers (vincent.cuypers@uhasselt.be) is a doctoral student in biology and philosophy of science at Hasselt University, in Diepenbeek, and at KU Leuven, in Leuven, Belgium. Frank E. Zachos (frank.zachos@NHM-WIEN.AC.AT) is an evolutionary zoologist with the Natural History Museum Vienna, in Vienna, Austria, and with the Department of Genetics at the University of the Free State in Bloemfontein, South Africa. Andreas De Block (andreas.deblock@kuleuven.be) is a philosopher of science at KU Leuven, in Leuven, Belgium.

Abstract

When producing species classifications, taxonomists are often confronted with gray-area cases. For example, because of incipient or shallow divergence, it can be scientifically valid both to split groups of organisms into separate species and to lump them together into one species. It has been claimed that, in such cases, the ranking decision is, in part, subjective and may differ between taxonomists because of differences in their conceptions of species or even in conservation values. In the present article, we use a vignette study to empirically test this claim and to explore the drivers of taxonomic decision-making in gray-area cases. For three fictional taxonomic scenarios, we asked the opinion of a sample of taxonomists on one of slightly different versions of an abstract containing a decision on species status. The cases were explicitly designed to represent gray-area cases, and the differences between versions related to potential drivers of decisions, such as information on conservation status, different kinds of additional evidence, and information on the presence or absence of gene flow. In general, our results suggest that taxonomists tend to disagree at least moderately about species-ranking decisions in gray-area cases even when they are presented with the same data. We did not find evidence that species concepts or conservation values are strong drivers of taxonomic disagreement. Instead, operational concerns, such as the presence or absence of different kinds of data, seemed to be more important.

Keywords: gray-area taxa, species delimitation, methods in taxonomy, vignette study, taxonomic disagreement

For many taxa across the Tree of Life, specialists in taxonomy disagree about how to classify them. Such disagreements often revolve around the rank of groups—for example, whether they should be recognized as species or as subspecies. This is often explained by the fact that taxonomic classifications enforce a binary system—a group of organisms either is recognized as a species or is not—onto differences that are usually gradual and continuous rather than discrete (Zachos et al. 2020, Thiele et al. 2021). Many of the criteria in use for delimiting species, such as morphological or molecular distinctness, interfertility, or ecological niche differentiation, indeed apply to groups of organisms in various degrees. Although many groups are clearly distinct, probably warranting species status, other groups find themselves in a gray area between what are typically accepted as good separate species and what are not. The appropriate ranking decision in such gray-area cases is not clear cut, and taxonomists may disagree even if they use the same data and criteria.

Because ranking decisions in gray-area cases often diverge among taxonomists, some have called ranking decisions in taxonomy at least partly subjective (Mishler and Wilkins 2018, Zachos et al. 2020, Zachos 2022). If taxonomy is indeed subjective in that way, that could pose problems for the discipline. Not only could it fuel unnecessary debates in a discipline that already lacks funding and researchers, but it would also affect all scientific and nonscientific domains that rely on the species-level classifications that taxonomists generate (see, e.g., Faurby et al. 2016, Willis 2017, Cuypers et al. 2022). These domains typically assume that all units at the species level are similar and, importantly, directly comparable, but, if ranking decisions are sometimes executive decisions rather than completely evidence based, this assumption may be unfounded. Disagreements often lead to the circulation of competing classifications (McClure et al. 2020, Neate-Clegg et al. 2021), which results in different groups of users using different classifications—making synergizing efforts often difficult—and forces users to invest in taxonomic decision-making themselves.

In order to reduce the disorder that disagreements and uncertainty create in taxonomy, it is important to know what drives taxonomists' decisions and disagreements in gray-area cases. Three main factors are regularly cited as causing taxonomists to disagree about the appropriate taxonomic treatment of a particular taxon. First, it is commonly claimed that taxonomists' preferred species concept could explain why their conclusions differ from those of colleagues. In particular, it is often assumed that taxonomists preferring the Biological Species Concept (BSC) are less likely to split taxa into smaller groups than are taxonomists who prefer the diagnosability version of the Phylogenetic Species Concept (dPSC) (Agapow et al. 2004, Isaac et al. 2004).

Second, some have argued that taxonomic disagreement about gray-area cases is driven by differences in the way species concepts are methodologically operationalized (Camargo and Sites 2013, Satler et al. 2013, Conix 2018). In the case of the BSC and the diagnosability version of the dPSC, this translates into debates about the importance of gene flow and (cryptic) molecular

Received: April 28, 2023. **Revised:** August 17, 2023. **Accepted:** August 30, 2023

differentiation and about the various ways in which an abstract notion such as gene flow can be tested in practice.

Finally, and more controversially, some have claimed that taxonomists are sometimes influenced by nontaxonomic considerations, such as the implications of ranking decisions for conservation (Karl and Bowen 1999, Isaac et al. 2004). In that argument, the claim is usually that taxonomists are more likely to recognize threatened groups as distinct taxa (species or subspecies), hoping that this would improve the chances of legal protection or conservation action for those groups. Other value-laden factors that potentially play a role are economic, political, or sociological. It is commonly believed, for example, that there were strong lumping traditions in both bird and mammal taxonomy in the past (Cotterill et al. 2014, Sangster 2014). Similarly, more experienced taxonomists and those working in low-income countries may rely more on morphological evidence (as an operationalization) than young taxonomists and those working in high-income countries do. Taxonomists working in countries with relatively low diversity, on the other hand, may have a stronger tendency to split (Harris and Froufe 2005).

To our knowledge, these explanations have never been experimentally tested. Although there have been at least two surveys on which species concepts biologists use and how they use them (Pušić et al. 2017, Stankowski and Ravinet 2021), these were simple self-report surveys among biologists of many subdisciplines. This study, instead, is experimental and only included responses from practicing taxonomists. Our aims are to test whether taxonomists indeed sometimes make different ranking decisions given the same data (subjectivity) and to investigate what kinds of taxonomic and nontaxonomic information—particularly, species concepts, evidence types (operationalization), and conservation values—are most likely to influence ranking decisions. To accomplish this, we carried out an online vignette study in which respondents were asked to evaluate three fictional taxonomic cases. For each case, any single respondent was presented with one of multiple slightly differing versions of the same abstract and had to state whether they agreed with the decisions made in that abstract. This allowed us to quantify variation in the responses of taxonomists in general and variation between groups of taxonomists that had received a different version of the abstracts.

Methods

This study was approved by the Social and Societal Ethics Committee of KU Leuven, Belgium (file no. G-2022-4955-R2(MIN)). Apart from the country of residence, no personal data were collected, and the data were published with the country of residence aggregated into continents and low- or high-income country (as named and classified by the World Bank 2022) to guarantee the anonymity of the respondents. Data collection only started after the full research design was preregistered on the Open Science Framework. The full questionnaire, analysis plans, raw data, analysis code, and supplemental materials can be found on the Open Science Framework (OSF) page (Conix et al. 2022) of the research project.

Design

We designed an online survey consisting of questions about the respondents' characteristics and fictional taxonomic abstracts (vignettes). The questionnaire was designed by the authors of this study, and revised after feedback from working taxonomists and a pilot with 14 taxonomists. The respondent characteristics in the final version included whether the respondent is a taxonomist, whether they do this professionally, their experience, their country of residence, their taxon of specialization, whether they read taxonomic literature outside of their area of expertise, and their preferred species concept. To avoid influencing the responses to the vignettes, the respondents were asked about their preferred species concept only after evaluating the vignettes.

Table 1. All cases and the number of respondents for each case and condition.

Case	Condition	N	*n* (hypothesis test)
Plant	Neutral	143	127
	Threatened	151	134
	Abundant	143	128
Frog	Neutral	119	103
	Morphology	107	98
	Mitochondrial DNA	105	90
	Ecology	110	100
Flatworm	Neutral	155	134
	Gene flow	139	128
	No gene flow	141	124

Note: Because we collected additional data after the preregistered period, the number of participants for the hypothesis tests differs from the total number of participants.

Each participant was given three vignettes in random order. These included one abstract describing a new fictional plant species, one abstract describing a new fictional frog species, and one abstract describing a new fictional flatworm species. All three fictional taxa were designed to be gray-area cases. We chose to use fictional taxa to avoid the possibility that taxonomists' preexisting opinions on real taxa would influence their decision. We chose a plant, frog, and flatworm in order to have at least one taxon that the respondents would be likely to know little about (the flatworm), one taxon that the respondents are likely to be somewhat familiar with (the frog), and one nonanimal case (the plant).

For each vignette, there were several versions that we designed to differ as little as possible, apart from the condition under investigation. The plant case was designed to investigate the role of conservation values. We included a version of the vignette stating that the taxon is threatened, a version stating that the taxon is not threatened, and a neutral version with no information about the conservation status. The frog case was designed to test the role of operationalization, and the versions differed in the kinds of evidence types they included. The neutral version included only limited morphological data. The other versions added more morphological data, mitochondrial DNA data and ecological data, respectively. The flatworm case centered on gene flow, with a version mentioning gene flow, a version mentioning the absence of gene flow, and a neutral version not mentioning gene flow at all. Because gene flow is tightly related to reproductive isolation, this vignette served as a test of the influence of species concepts on ranking decisions. See table 1 for an overview of all cases and conditions.

Each respondent was randomly assigned one version of each of the three cases. For each case, the respondents were asked whether they agreed with the ranking decision (i.e., with the proposed new species) in the abstract. This is the main outcome variable of the study. For the frog case, they were also asked which kind of evidence they thought was lacking in case they did not agree with the ranking decision in the abstract. Each respondent was also asked whether they would accept the

abstract for a conference presentation. This question was included to check whether the respondents perceived the abstracts as scientifically legitimate. The full survey, with vignettes, is available in supplemental document S1.

Sampling

The survey was distributed through several taxonomic mailing lists, with one reminder 2 weeks after sending out the survey. In addition, the survey was disseminated through the networks of the authors and sent to various professional organizations and natural history museums asking them to disseminate the survey (for a full list of organizations and institutions contacted, see supplemental document S2). Because the number of responses from taxonomists residing outside of Europe was low after the full preregistered sampling period, we kept the survey open for 1 month longer than initially planned and sent the survey out through our networks in South America, Africa, and Southeast Asia. Because we expect the sample to be more representative of the population after the additional sampling effort, all exploratory analyses reported below were done using the extended data set. Because the additional sampling was not preregistered, the registered hypothesis tests were done using the original, smaller data set.

In order to use only high-quality data for the analysis, we only retained responses that took at least 150 seconds to finish the survey (i.e., the minimum time needed to read and process all questions) and responses that included a reply to at least one of the three main outcome questions. We also only retained responses from the respondents who indicated that they were taxonomists.

It is likely that some participants received the invitation for the survey more than once because of overlaps between different channels of dissemination. Because of the snowball method of sampling, it is not possible to estimate the response rate and difficult to estimate how representative the sample is of the wider population.

Statistical analysis

We registered two hypotheses for this study on OSF: First, if taxonomists frequently make different ranking decisions even when given the same data, then there will be strong disagreement about the ranking decision for the three abstracts across conditions. Second, if taxonomists are more likely to rank a group as a species if it is threatened, then the proportion of *agree* responses will be substantially higher for the participants assigned to the threatened condition than for the participants assigned to the not threatened or neutral condition.

The aim of the first hypothesis was to establish that taxonomists indeed make divergent ranking decisions, even when given the same data. Although this may be obvious to most working taxonomists (see, e.g., Isaac et al. 2004, Tattersall 2007, Heller et al. 2013), it has, to our knowledge, never been experimentally quantified. We tested this hypothesis using a Bayesian model to estimate the proportions of *agree* responses for all conditions for each of the three cases. We decided in advance that we would consider there to be strong disagreement about an abstract when the entire 80% highest density interval (hdi; i.e., the smallest 80% credible interval) of the estimated proportion of the minority opinion for that abstract was above .25. This would mean that it is highly likely that at least 25% of taxonomists would have a different opinion about the ranking decision in that abstract than the majority opinion.

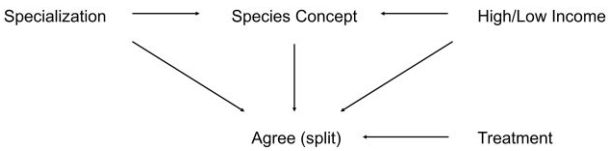

Figure 1. Directed acyclic graph showing the causal model of ranking decisions assumed in the exploratory analysis. ***Agree (split)*** is the outcome variable, and ***treatment*** are the various versions of the vignettes for each of the three cases.

The aim of the second hypothesis was to test whether, as some have claimed (Isaac et al. 2004, Conix 2019), nontaxonomic considerations such as conservation values influence ranking decisions. To test this hypothesis, we estimated the proportion of *agree* responses for the threatened and abundant conditions in the plant case using a Bayesian model and subtracted the posterior distributions of these proportions. We stipulated in advance that we would accept the hypothesis if 0 fell outside of the 80% hdi of the resulting distribution (meaning that the estimated difference between the two conditions is highly unlikely to be 0).

In addition to these two registered hypothesis tests, we designed a causal model incorporating the main factors cited in the literature as potential causes of disagreement. This causal model, which is represented in figure 1 (Cinelli et al. 2022), is based on our experience of the field and the literature on the species problem and could serve as a basis for future research into the causes of ranking disagreement. Note that this model only captures what we are assuming to be the main factors and omits other factors that we did not see a clear causal role for but that may still have an influence, such as whether the taxonomist works on living or fossil taxa, their seniority, and their gender. Future research could expand this model in case there is support for other factors playing a role. As it is, the model shows that we assume that ranking decisions are influenced by the species concept of taxonomists and the particulars of the case (i.e., the treatments) but also by which taxon they specialize in and whether they reside in a low-income country. The latter two factors capture, among other things, the research community and research culture taxonomists are active in and might influence ranking decisions both directly and through an influence on their species concept. By including the region of residence (low-income versus high-income countries) in the model, we aim to avoid the causal estimates being affected by the bias toward participants in the sample who reside in North America and Europe.

We used this causal model to select predictors for regressions testing the causal role of conservation values in the plant case (model 1), operationalization in the frog case (model 2), and species concepts in the flatworm case (model 3). More precisely, we applied the so-called backdoor criterion to the causal model to select the variables to condition on and avoid the estimates being influenced by noncausal paths in the model (Cinelli et al. 2022). In addition, we tested the conditional independencies of this model where possible to ensure that there were no strong associations between variables where the model did not predict this to be the case (Cinelli et al. 2022). Note that these models were exploratory and designed after collecting and seeing the data. Therefore, even though they were designed using a causal model, they should be interpreted with caution and mostly as the basis for designing further—ideally, preregistered hypothesis-driven—research.

All three models were Bayesian logistic regressions with an *agree* response as the outcome, the treatment and species concept as the predictors of interest, and the taxon of specialization

Table 2. Statistical models for the exploratory analysis of ranking decisions.

Case	Outcome	Cause of interest	Implementation cause
Plant (model 1)	Agree	Values	Species concept, treatment (neutral, threatened, abundant)
Frog (model 2)	Agree	Operationalization	Species concept, treatment (neutral, morphology, DNA, habitat)
Flatworm (model 3)	Agree	Species concept	Species concept, treatment (gene flow, no gene flow, neutral)

Note: In all cases, income and specialization were controlled for.

and income status of the home country (high or low) as control variables. They all included a general intercept and offsets for the included groups (income, species concepts, taxon of specialization) and treatments (the conditions of the vignettes). In model 1, we also included a varying effect of treatment by income status, because we expected that the influence of conservation values might differ with respect to income level. Similarly, the effect of treatment varied by taxon of specialization and income status in model 2, because we assumed that the influence of different operationalizations could differ according to the taxonomists' specialization and whether they worked in a low- or high-income country. Finally, we included a varying effect for species concept by treatment in model 3, because we expected that the influence of the gene-flow condition might differ depending on the species concept participants subscribed to.

Only participants who responded to all questions included in the analysis (agreement, income status, taxon of specialization, species concept) were included in the analysis ($N = 423$). We used Bayesian models for all analyses because their results are more intuitive to interpret than the outcomes of traditional frequentist methods (Kruschke 2013), because Bayesian models more easily enable pooling information within groups (such as specializations, region of residence or levels of seniority) through a hierarchical structure (McElreath 2016), and because they are well suited for analyses in which some groups have small sample sizes. We used weakly informative priors in all these regressions, and all the analyses were accomplished using Markov chain Monte Carlo methods (van Ravenzwaaij et al. 2018). For a full specification of the models, as well as the Pymc code used to run them, see the analysis code on the OSF page of the project (https://osf.io/qbmea). For an overview of the three models, see table 2. We used Pandas (McKinney 2010), Scipy (Virtanen et al. 2020), Numpy (Harris et al. 2020), Seaborn (Waskom et al. 2022), and Matplotlib (Hunter 2007) in a Jupyter Notebook for all descriptive analyses. We used the Pymc (Salvatier et al. 2016), Bambi (Capretto et al. 2022), and ArviZ (Kumar et al. 2019) libraries in Python (see the source code for all package versions) for all hypothesis tests and exploratory regressions.

Results and discussion

After both sampling periods, the survey was filled in by 706 participants. After removing responses that took less than 150 seconds, responses without answers to the main outcome questions ("Do you agree?"), and the responses of participants that indicated they were not taxonomists (97 in total), 447 responses were left. This is substantially more than in previous surveys on species concepts (Pušić et al. 2017, Stankowski and Ravinet 2021)—in particular, if only taxonomists are considered. For the two hypothesis tests, we removed the responses received after the preregistered sampling period ($n = 51$), as well as responses with missing data for one of the variables included, keeping 396 responses for hypothesis 1 and 389 for hypothesis 2.

The main respondent characteristics for this data set of 447 respondents are summarized in tables 3a, 3b, and 4, and are visualized in supplemental figure S1. The respondents were relatively equally divided among various species concepts. The dPSC (28.3%) was most popular, closely followed by the BSC (24.5%) and the Evolutionary Species Concept (24.1%). It is notable that 43.6% of our sample resided in Europe, and only 27.9% in Asia, Africa, or South America. Given the strong increase of taxonomists particularly in South America and the Asian-Pacific region over the past decades (de Carvalho et al. 2005, Costello et al. 2013), this indicates that our sample was still biased toward taxonomists residing in Europe and North America. This is probably because the survey was only distributed in English and through our own networks and, therefore, still had lower visibility in these regions, despite our additional sampling efforts. Over 30% of our sample had worked for at least 30 years since their PhD. This may be partially due to the bias toward North America and Europe in the sample, because many of the taxonomists who reside there might be close to retirement (but see Costello et al. 2013). Not surprisingly, the distribution of specializations was, in most cases (but not always), clearly different from the proportion of the Tree of Life that the specialization takes up. In particular, specialists in reptiles and amphibians, plants, mammals, and birds made up 30.1% of the sample, even though the taxa they are specialized in make up only a relatively small fraction of all species.

The demographic makeup of the sample is in line with that of another relatively large survey among taxonomists (Salvador et al. 2022) and our expectation that taxonomists in senior positions and taxonomists working on vertebrates (and insects) make up a relatively large share of all taxonomists. The distribution of specializations is also broadly in line with the proportions of mentions of taxa from these groups in a large full-text corpus of taxonomic research papers (https://philarchive.org/rec/PENMCA-2 [preprint: not peer reviewed]). However, like our survey, this survey and the full-text corpus use convenience samples. Given the method of sampling that was used, it is unlikely that these samples are representative, because certain demographics may be more likely to participate in the survey than others. Because we have no clear hypotheses or information about what the potential sources of bias may be, we did not include them in the statistical models. Hence, we urge the readers to interpret the results of our analyses with caution, because they may be biased by our sampling method.

Disagreement

The responses (agree or disagree) for all conditions for all cases are summarized in figure 2. For each condition, the participants were more likely to accept the abstract for a conference than they were to agree with the ranking decision (see supplemental figure S6). This suggests that the respondents did not interpret the agreement question as one about scientific quality and that the abstracts were generally seen as academically acceptable.

Table 3a. Years since the start of taxonomic activity and continent of residence of the respondents.

		Percentage of respondents per seniority group				
Continent	***n***	**0 to 5 years**	**6 to 10 years**	**11 to 20 years**	**21 to 30 years**	**More than 31 years**
Africa	17	17.6	17.6	23.5	5.9	35.3
Asia	50	8	22	44	12	14
Europe	191	8.4	13.6	26.2	26.2	25.7
North America	100	5	6	17	25	47
Oceania	27	3.7	3.7	7.4	14.8	70.4
South America	56	5.4	12.5	30.4	21.4	30.4

Table 3b. Years since the start of taxonomic activity and species concept of the respondents.

		Percentage of respondents per seniority group				
Species concept	***n***	**0 to 5 years**	**6 to 10 years**	**11 to 20 years**	**21 to 30 years**	**More than 31 years**
Biological Species Concept	104	5.8	11.5	18.3	25	39.4
Evolutionary Species Concept	103	10.7	13.6	33	19.4	23.3
Genetic Cluster Species Concept	12	0	0	0	41.7	58.3
Phylogenetic Species Concept, diagnosability version	120	5.8	10.8	31.7	20.8	30.8
Phylogenetic Species Concept, monophyly version	42	9.5	14.3	23.8	11.9	40.5
Other	45	6.7	13.3	15.6	28.9	35.6

Table 4. Taxon specializations and income status of the country of residence of the respondents.

Taxon	***n***	**Low income (as a percentage)**	**High income (as a percentage)**
Algae	6	16.7	83.3
Birds	11	36.4	54.5
Fishes	28	14.3	85.7
Fungi	5	0	100
Insects	135	34.1	65.9
Mammals	30	36.7	63.3
Molluscs	19	21.1	73.7
Noninsect arthropods	61	32.8	65.6
Nonvertebrate deuterostomes	4	50	50
Plants	59	32.2	67.8
Prokaryotes	3	0	100
Protists (nonalgae)	7	42.9	57.1
Remaining invertebrates	46	26.1	71.7
Reptiles and amphibians	33	45.5	54.5

The estimated proportion of *agree* responses for each condition for each case is listed in table 5. Disagreement (the size of the minority opinion) within conditions was above .25 for the entire 80% hdi for 4 out of 10 conditions (combining the three cases). Therefore, our hypothesis that there would be strong disagreement about gray-area cases was not confirmed according to the criteria we had selected. This was due in particular to the plant case, which had relatively high levels of agreement. Still, all conditions showed at least moderate disagreement, with an average proportion of 27.84% for the minority opinion across all conditions (see supplemental figure S2) and disagreement means ranging between 17.7% and 45.9%.

Of course, these results should be interpreted with caution. On one hand, they might underestimate the true rates of disagreement if the plant case was too clear cut and did not represent a true gray-area case. More generally, there were substantial differences in disagreement between conditions and cases, with high levels of agreement for all three plant conditions and the mitochondrial DNA condition of the frog case. This suggests that levels of disagreement are sometimes very case dependent, and it remains an open question to what extent we can generalize findings about them to other cases.

On the other hand, this study might overestimate disagreement as well. First, the vignettes were explicitly designed to be gray-area cases that are likely to elicit disagreement. This means that the results only apply to such cases and not across the whole hierarchical realm covered by taxonomy (in line with findings by Faurby et al. 2016). Many cases of species delimitation will be uncontentious. Second, the vignettes in this study were short abstracts, and the participants were asked to evaluate the abstracts even if they were outside their taxon of specialization. This is unlike taxonomic reality, in which ranking decisions are typically not made on the basis of information that can be given in an abstract of 150 words, and taxonomists rarely have to make ranking decisions outside their taxon of expertise. Therefore, it may also be that part of the disagreement in these cases was caused by the lack of information in the abstracts. This is suggested by the fact that, in the frog case, disagreement decreased with more information (i.e., going from neutral to one of the conditions with extra evidence). However, although it is true that the vignettes provided less information than taxonomists typically work with, it should be remembered that working with little information is the reality of many taxonomic decisions. Therefore, although the lack of information might exacerbate the disagreement in the present study, we do not think it is merely a product of our methods.

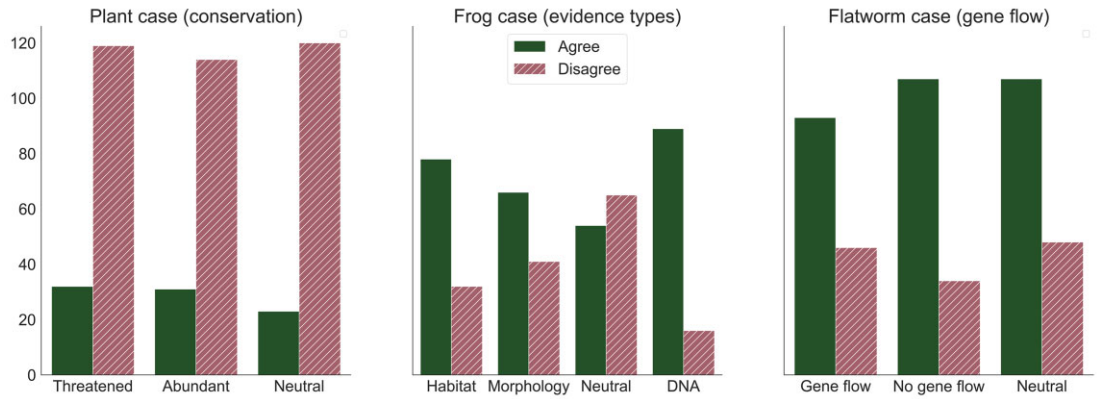

Figure 2. Responses to the question "Do you agree with the ranking decision in the abstract?" for each of the conditions of each of the cases for the full data set. ***Agree*** indicates agreement with recognizing a new species rather than lumping it.

Table 5. Estimated proportion of ***agree*** for each condition for each of the three cases.

		Hypothesis test results					Exploratory models results		
Category	Condition	Mean	Standard deviation	10% hdi	90% hdi	Minority more than .25	Mean	10% hdi	90% hdi
Plant	Neutral	.177	.033	.131	.216	No	.135	.084	.182
	Abundant	.221	.036	.173	.265	No	.162	.101	.212
	Threatened	.219	.035	.17	.261	No	.232	.167	.292
Frog	Neutral	.459	.048	.395	.518	Yes	.450	.380	.524
	DNA	.82	.04	.772	.874	No	.863	.818	.916
	Habitat	.731	.043	.677	.788	No	.701	.634	.774
	Morphology	.609	.048	.548	.672	Yes	.611	.538	.686
Flatworm	Neutral	.696	.039	.648	.748	Yes	.666	.605	.736
	Gene flow	.659	.041	.608	.713	Yes	.675	.608	.745
	No gene flow	.758	.038	.712	.809	No	.765	.712	.826

Abbreviation: hdi, highest density interval.

Drivers of disagreement: Species concepts and operationalization

Supplemental tables S2, S3, and S4 list the coefficients and highest density intervals for the selected variables of interest for models 1–3 (the full results are reported in table S3 and are not reported in the article to avoid the table 2 fallacy; Westreich and Greenland 2013). In all three models, the influence of species concepts on accepting the species descriptions was close to zero. As we expected, the effect of species concepts was strongest in the flatworm case, which was centered on gene flow (figure 3). Because reproductive isolation is the main criterion for species status in the BSC, we expected that the difference in expected proportions of *agree* responses for the gene flow and no gene flow conditions would be largest for proponents of the BSC: They should accept the species if there is no gene flow and should reject it if there is gene flow. However, not only was there a substantial group of proponents of the BSC that accepted the species even under the gene flow condition (mean expected proportion of .63), posterior predictive sampling from model 3 (artificially limiting the population in turn to the various combinations of treatments and species concepts) also showed that we should expect almost no difference between the different species concepts in how important gene flow is (figure 4). That is, the proportion of *agree* responses between gene flow and no gene flow was nearly identical across species concepts. More generally, the levels of disagreement within species concepts were very similar to the levels of disagreement across species concepts (supplemental figure S5). All of this suggests that the influence of species concepts on ranking decisions was small, and, if there was any, not directly related to the content of those concepts.

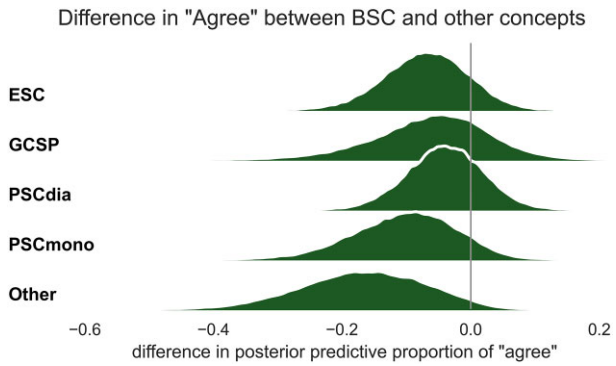

Figure 3. Density plots of the difference between the expected proportion of ***agree*** for the BSC and for other species concepts using posterior predictive samples from model 3 (the flatworm case). A value above zero indicates a higher tendency to accept the new species for the BSC than for the concept in question. These posterior predictive samples were drawn from model 3, leaving the demographic characteristics of the sample intact but in turn changing the species concept to each of the included concepts for the entire sample.

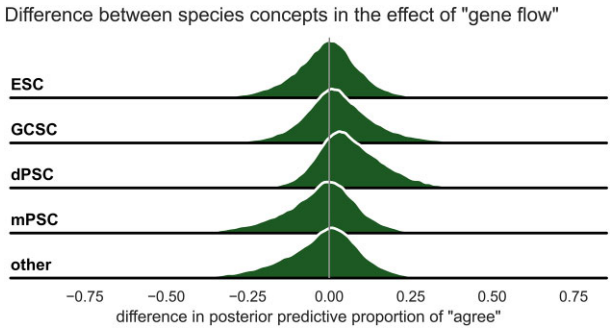

Figure 4. Density plots of how the BSC and each of the other species concepts differ in the difference of proportion of "gene flow" and "no gene flow" in posterior predictive samples drawn from model 3 (the flatworm case; drawn for each combination of treatments and species concepts, keeping the other demographic properties of the sample intact). A value above zero indicates that the presence or absence of gene flow tended to make a bigger difference for proponents of the BSC than for the concept in question.

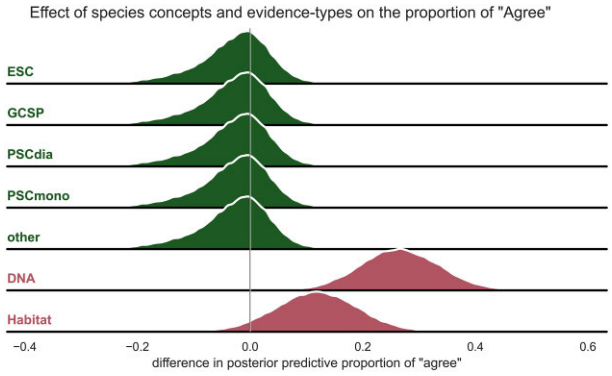

Figure 5. Density plots of the expected differences in proportion of ***agree*** between the BSC and other concepts, and between morphology and other treatments. The expected proportions for the treatments and species concepts were generated using posterior predictive samples from model 2 (the frog case), keeping the other demographic properties of the sample intact.

Contrary to species concepts, operationalization did seem to have a strong influence on agreeing with the ranking decision in the abstract. The model for the frog case, which was designed to test the influence of operationalization, shows that the rates of disagreement differed substantially between the treatments. In particular, the evidence of mitochondrial DNA differentiation appeared to be a far stronger reason to recognize the frog as a species than was morphological and ecological evidence. Figure 5 shows that although the posterior predictive proportions of *agree* responses hardly differed among species concepts, they differed strongly between morphology on one hand and mitochondrial DNA and habitat on the other hand. This shows that, for the frog abstract, operationalization was far more influential than were species concepts. The difference between morphological evidence and the other operationalizations also differed between groups, with taxonomists working in low-income countries accepting it more often as sufficient for species status (figure 6). We suspect this may be the case because taxonomists in low-income countries either do not always have the resources to produce molecular evidence or are more likely than taxonomists in high-income countries to consider molecular evidence as just one among many useful tools to use, rather than the single main piece of evidence to rely on.

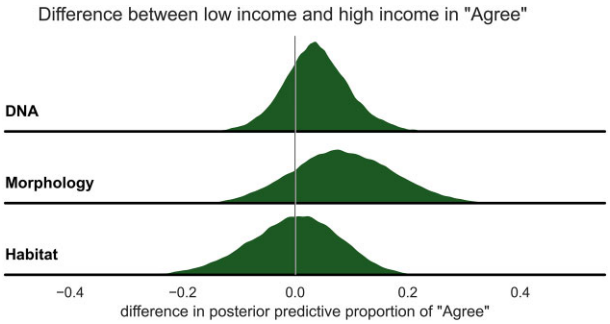

Figure 6. Density plots of the expected differences in the proportion of ***agree*** between high- and low-income countries for the difference between the neutral treatment and other treatments. The expected proportions for the different treatments were generated using posterior predictive samples from model 2, keeping the other demographic properties of the sample intact.

Drivers of disagreement: Conservation values

We found no difference (mean = .003, 80% hdi = –.061, .057; see supplemental figure S8) in the estimated proportion of *agree* responses of the threatened and abundant versions of the plant case for the sample of the hypothesis tests. This suggests that, in this case, conservation status did not influence ranking decisions and that our second hypothesis, concerning a role for conservation values in taxonomic decision-making, is therefore disconfirmed. Although it may be that other cases would have shown an effect, it is at least tentative evidence against the commonly made claim in the literature that taxonomists sometimes tend to recognize threatened groups as species merely to improve their chances of getting funding for conservation action (Isaac et al. 2004, Conix 2019). It should be noted, however, that there was less disagreement in general about the plant case than about the other two cases. Therefore, as we already mentioned, an alternative explanation may be that the vignette was not considered a gray-area case by the respondents and, because of that, did not show conservation values to play a role.

It should also be noted that, for the extended sample, with additional sampling effort, model 1 expects the proportion of *agree* responses to be higher for groups with the threatened version of the vignette if we take posterior predictive samples from the model assuming the demographics of our study population. This is not due to differences in the coefficients for the threatened and abundant groups (see supplemental figure S7) but is due to the fact that the model finds a clear difference between threatened and abundant for taxonomists working in low-income countries (which we tried to sample from in the second round of sampling, excluded from the hypothesis test; see figure 7). We speculate that this may be the case because the fictional plant case was set in a tropical environment. Taxonomists from low-income countries are more likely to be active in such environments on a daily basis and may therefore have felt a stronger connection to the plant group in question. Another possibility is that taxonomists from the Global South are more concretely aware of the ongoing extinction crisis because the tropics are such a major stronghold of the world's biodiversity.

It is important to highlight the limitations of the plant case as a test of the role of nontaxonomic values in ranking decisions. For one, we only looked at conservation values. Other sociological factors may well play a role that may not have been captured by the plant case. For example, in all three models, the taxon of specialization, as well as the income status of the country of activity,

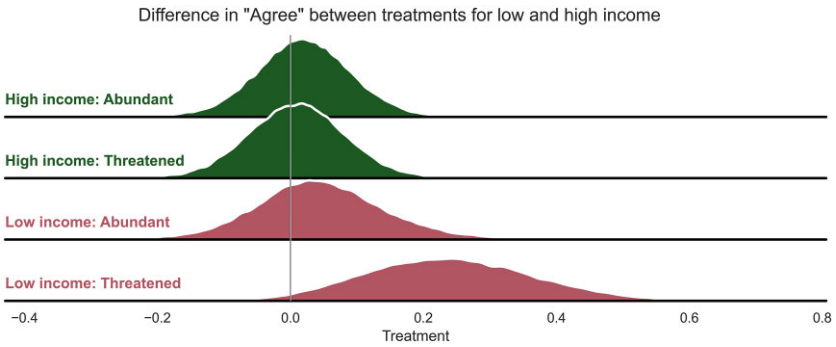

Figure 7. The difference in proportion of ***agree*** between the neutral and the two other treatments for posterior predictive samples from model 1 with the entire sample set to high income and low income. This shows that participants working in low-income countries were more likely to agree with the ranking decision in the threatened condition.

seemed to affect the tendency to agree with the ranking decision, with insect taxonomists and taxonomists from low-income countries showing a tendency to split. Therefore, it may well be that sociological factors such as country of residence and training, as well as varying academic culture and traditions in different taxonomic communities, are nontaxonomic considerations that influence ranking decisions. However, even if the causal model we assume (figure 1) implies that the coefficients for these variables are meaningful as an indication of direct (not total) causal effect, it is not possible with our data to draw definitive conclusions about this, and hypothesis-driven follow-up research is needed to confirm and flesh out these patterns. Another limitation is that the responses might have been different if, instead of a plant case, we had used a highly charismatic animal to test the influence of conservation values. Follow-up research focusing on the role of conservation values could take up our results and vary more features of the abstract, such as the kind of taxon that is used, the habitat in which it lives, or where it is found.

Conclusions and prospects

This survey indicates that there was at least moderate taxonomic disagreement about the fictional gray-area cases we presented. Even though taxonomists were given the same information about a group of organisms, there was an average of 28% disagreement about the groups' status as a separate species. Equally important is the light that our survey sheds on the drivers of this disagreement. Unlike what many researchers seem to believe, differences in adherence to species concept do not appear to lead to more differences in the observed taxonomic decisions, and adherence to the same species concept does not lead to lower levels of disagreement. The concrete operationalization of species concepts seems far more important for explaining taxonomic disagreement. Because these operationalizations are not strictly tied to a single species concept, this indicates that the disagreement may often be practical (information and methods) rather than theoretical (concepts). Finally, contrary to our expectations, conservation values did not seem to motivate taxonomic decisions—at least, not in general. Again, this contrasts with the attention that is given to the role of values for taxonomic decision-making in both the philosophical and biological literature (Isaac et al. 2004, Ludwig 2016, Conix 2019).

We draw two main concrete conclusions from this. First, our results suggest that, at least in gray-area cases, some degree of subjectiveness is sometimes probably hard to avoid if we insist on using the current Linnaean system, where taxa are given specific ranks: Although disagreement was not as high as we expected, there was at least moderate disagreement about every case. The fact that disagreement is probably most common in gray-area cases should not be taken to entail that more information on the groups will (always) solve disputes on species status. Although we did find in the frog case that more information might sometimes reduce disagreement, this is not the silver bullet some consider it to be. For one, evidence is sometimes lacking, and it is not always possible or feasible to gather more data. Moreover, there are also cases in which different lines of evidence conflict (Satler et al. 2013). Simply adding information is unlikely to solve all problems, because speciation is inevitably a multifaceted and gradual process. Collecting more data does not turn shallow divergence into deep divergence, and even with an abundance of information (dichotomous) decisions remain difficult in such cases.

This reality need not reflect negatively on taxonomy as a scientific discipline or on the work of taxonomists; there are parallels in equally respectable fields of science (Slater 2017, Cuypers and De Block 2023). In genuine gray-area cases, the uncertainty about species status and the resulting taxonomic disagreement refer to the ranking part of taxonomy and do not necessarily reflect ignorance about what a species is or about particular characteristics of the group under consideration. Rather, they are an inevitable consequence of the imposition of a binary system onto a nonbinary, continuous reality (Zachos et al. 2020, Thiele et al. 2021).

We do believe, however, that taxonomists should keep this reality of subjectivity in mind and take some measures to alleviate unwanted consequences. For example, we believe it is important that taxonomists provide full transparency on why they decide what they decide. Taxonomists should provide detailed methodological information (as our results show, operational choices matter) and information on how they interpret their results and translate them into taxonomic decisions. One step in this direction could be to register taxonomic methods and criteria for attributing species status in advance. As some of the authors of this study have argued elsewhere, the preregistration of research methods has beneficial effects on transparency and clarity in many disciplines and could also be of use in taxonomy (Conix et al. 2023). More generally, this subjectivity implies that we should not assume that groups at the same rank are always comparable or similar and that we should be very careful in using Linnaean ranks for inferential or practical purposes (see also Faurby et al. 2016, Willis 2017). This conclusion is in line with other arguments against the use of Linnaean ranks and provides support for alternative systems of rankless classification (Mishler and Wilkins 2018, Mishler 2021).

The degree of subjectivity involved in taxonomic decision-making in gray-area cases should also be acknowledged when assessing—at times vehement—taxonomic disagreements. If the disagreements concern empirical questions—for example, on evolutionary patterns in the groups under consideration—they obviously have scientific value. But if the disagreements turn out to be a pure matter of appreciation, it may not always be useful to pursue debates about them endlessly, given the urgent demands for clear and stable taxonomies in biology and beyond. Rather, it may be advisable to take recourse to procedures suited to arbitrate executive issues of that kind—for example, through some form of taxonomic list governance (Garnett et al. 2020). This is precisely what the four main global bird lists are currently doing, unifying their diverging lists through a voting procedure (McClure et al. 2020, Cuypers and De Block 2023).

The second main implication of our results is that a shift may be needed in what philosophers and biologists should focus on when they study the conceptual side of the species problem. Our results suggest that the research community should probably spend more time researching the role of operationalization in ranking decisions and should focus less on studying how species concepts and nonepistemic values may shape taxonomy. This dovetails nicely with the first implication, because what we need is renewed reflection on how to deal with gray-area cases in taxonomic practice.

Acknowledgments

Stijn Conix's work for this article was funded by the Fonds de la Recherche Scientifique—FNRS under grant no. T.0177.21. Vincent Cuypers's work for this article was funded by the Research Council Flanders (FWO) under grant no. G0D5720N.

Supplemental material

Supplemental data are available at *BIOSCI* online.

References cited

Agapow P-M, Bininda-Emonds ORP, Crandall KA, Gittleman JL, Mace GM, Marshall JC, Purvis A. 2004. The impact of species concept on biodiversity studies. *Quarterly Review of Biology* 79: 161–179.

Camargo A, Sites JJ. 2013. Species delimitation: A decade after the renaissance. Pages 225–247 in Pavlinov IY, ed. *The Species Problem: Ongoing Issues*. InTech.

Capretto T, Piho C, Kumar R, Westfall J, Yarkoni T, Martin OA. 2022. Bambi: A simple interface for fitting Bayesian linear models in Python. *Journal of Statistical Software* 103: 1–29.

Cinelli C, Forney A, Pearl J. 2022. A crash course in good and bad controls. *Sociological Methods and Research* 2022: 00491241221099552.

Conix S. 2018. Integrative taxonomy and the operationalization of evolutionary independence. *European Journal for Philosophy of Science* 8: 587–603.

Conix S. 2019. Taxonomy and conservation science: Interdependent and value-laden. *History and Philosophy of the Life Sciences* 41: 15.

Conix S, Block A, Cuypers V, Zachos F. 2022. Exploring the reasons of diverging views in taxonomy. Center for Open Science. https://osf.io/qbmea. doi:10.17605/OSF.IO/QBMEA

Conix S, Cuypers V, Zachos F, Artois T, Monnens M. 2023. A plea for preregistration in taxonomy. *Megataxa* 10: 1–14.

Costello MJ, May RM, Stork NE. 2013. Can we name Earth's species before they go extinct? *Science* 339: 413–416.

Cotterill FPD, Taylor PJ, Gippoliti S, Bishop JM, Groves CP. 2014. Why one century of phenetics is enough: Response to "Are there really twice as many bovid species as we thought?" *Systematic Biology* 63: 819–832.

Cuypers V, De Block A. 2023. Resolving conceptual conflicts through voting. *Foundations of Science* 2023: s10699-023-09903-2. https://doi.org/10.1007/s10699-023-09903-2.

Cuypers V, Reydon TAC, 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.

de Carvalho MR, et al. 2005. Revisiting the taxonomic impediment. *Science* 307: 353–353.

Faurby S, Eiserhardt WL, Svenning J-C. 2016. Strong effects of variation in taxonomic opinion on diversification analyses. *Methods in Ecology and Evolution* 7: 4–13.

Garnett ST, et al. 2020. Principles for creating a single authoritative list of the world's species. *PLOS Biology* 18: e3000736.

Harris J, Froufe E. 2005. Taxonomic inflation: Species concept or historical geopolitical bias? *Trends in Ecology and Evolution* 20: 6–7.

Harris CR, et al. 2020. Array programming with NumPy. *Nature* 585: 357–362.

Heller R, Frandsen P, Lorenzen ED, Siegismund HR. 2013. Are there really twice as many bovid species as we thought? *Systematic Biology* 62: 490–493.

Hunter JD. 2007. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* 9: 90–95.

Isaac NJB, Mallet J, Mace GM. 2004. Taxonomic inflation: Its influence on macroecology and conservation. *Trends in Ecology and Evolution* 19: 464–469.

Karl SA, Bowen BW. 1999. Evolutionary significant units versus geopolitical taxonomy: Molecular systematics of an endangered sea turtle (genus Chelonia). *Conservation Biology* 13: 990–999.

Kruschke JK. 2013. Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology General* 142: 573–603.

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: 1143.

Ludwig D. 2016. Ontological choices and the value-free ideal. *Erkenntnis* 6: 1253–1272.

McClure CJW, et al. 2020. Towards reconciliation of the four world bird lists: Hotspots of disagreement in taxonomy of raptors. *Proceedings of the Royal Society B* 287: 20200683.

McElreath R. 2016. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.

McKinney W. 2010. Data structures for statistical computing in Python. Pages 56–61 in Jones E Millman J, eds. *Proceedings of the 9th Python in Science Conference (SciPy 2010)*. SciPy.org.

Mishler BD. 2021. *What, If Anything, Are Species?* CRC Press.

Mishler BD, Wilkins JS. 2018. The Hunting of the SNaRC: A snarky solution to the species problem. *Philosophy, Theory, and Practice in Biology* 10: 20180404. doi:10.3998/ptpbio.16039257.0010.001

Neate-Clegg MHC, Blount JD, Şekercioğlu ÇH. 2021. Ecological and biogeographical predictors of taxonomic discord across the world's birds. *Global Ecology and Biogeography* 30: 1258–1270.

Pušić B, Gregorić P, Franjević D. 2017. What do biologists make of the species problem? *Acta Biotheoretica* 65: 179–209.

Salvador RB, Cavallari DC, Rands D, Tomotani BM. 2022. Publication practice in taxonomy: Global inequalities and potential bias against negative results. *PLOS ONE* 17: e0269246.

Salvatier J, Wiecki TV, Fonnesbeck C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* 2: e55.

Sangster G. 2014. The application of species criteria in avian taxonomy and its implications for the debate over species concepts: Application of species criteria in practice. *Biological Reviews* 89: 199–214.

Satler JD, Carstens BC, Hedin M. 2013. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, Aliatypus). *Systematic Biology* 62: 805–823.

Slater MH. 2017. Pluto and the platypus: An odd ball and an odd duck: On classificatory norms. *Studies in History and Philosophy of Science A* 61: 1–10.

Stankowski S, Ravinet M. 2021. Quantifying the use of species concepts. *Current Biology* 31: R428–R429.

Tattersall I. 2007. Madagascar's lemurs: Cryptic diversity or taxonomic inflation? *Evolutionary Anthropology: Issues, News, and Reviews* 16: 12–23.

Thiele KR, et al. 2021. Towards a global list of accepted species I. Why taxonomists sometimes disagree, and why this matters. *Organisms Diversity and Evolution* 21: 615–622.

van Ravenzwaaij D, Cassey P, Brown SD. 2018. A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin and Review* 25: 143–154.

Virtanen P, et al. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* 17: 261–272.

Waskom M, et al. 2022. mwaskom/seaborn: v0.12.2 (December 2022). Zenodo. https://zenodo.org/record/7495530.

Westreich D, Greenland S. 2013. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology* 177: 292–298.

Willis SC. 2017. One species or four? Yes!…and, no. Or, arbitrary assignment of lineages to species obscures the diversification processes of neotropical fishes. *PLOS ONE* 12: e0172349.

World Bank. 2022. Low income. World Bank. https://data.worldbank.org/country/XM.

Zachos FE. 2022. *Critique of Taxonomic Reason(ing): Nature's Joints in Light of an "Honest" Species Concept and Kurt Hübner's Historistic Philosophy of Science. Species Problems and beyond*. CRC Press.

Zachos FE, Christidis L, Garnett ST. 2020. Mammalian species and the twofold nature of taxonomy: A comment on Taylor et al. 2019. *Mammalia* 84: 1–5.

Received: April 28, 2023. **Revised:** August 17, 2023. **Accepted:** August 30, 2023