Logistical and preference bias in participatory science butterfly data

Benjamin R Goldstein1,2, Sara Stoudt3, Jayme MM Lewthwaite4, Vaughn Shirey4,5, Eros Mendoza4, and Laura Melissa Guzman4*

The volume of and interest in unstructured participatory science data has increased dramatically in recent years. However, unstructured participatory science data contain taxonomic biases—encounters with some species are more likely to be reported than encounters with others. Taxonomic biases are driven by human preferences for different species and by logistical factors that make observing certain species challenging. We investigated taxonomic bias in reports of butterflies by characterizing differences between a dedicated participatory semi-structured dataset, eButterfly, and a popular unstructured dataset, iNaturalist, in spatio-temporally explicit models. Across 194 butterfly species, we found that 53 species were overreported and 34 species were underreported in opportunistic data. Ease of identification and feature diversity were significantly associated with overreporting in opportunistic data, and strong patterns in overreporting by family were also detected. Quantifying taxonomic biases not only helps us understand how humans engage with nature but also is necessary to generate robust inference from unstructured participatory data.

Data collected in participatory science programs, where volunteers report their encounters with wild organisms, are generated more rapidly, in greater volume, and with better spatiotemporal coverage and resolution than expert-generated survey data (Sullivan et al. 2014; Theobald et al. 2015; Di Cecco et al. 2021; Johnston et al. 2022). Because of these advantages, ecologists are increasingly turning to participatory science programs to support large-scale, high-resolution biodiversity monitoring, using statistical models to estimate species distributions, assess conservation priorities, identify population trends, and characterize species phylogeny (Theobald et al. 2015; Johnston et al. 2022). However, not all participatory data are equally appropriate for use in biodiversity monitoring. Some participatory science data are “semi-structured”, meaning they are associated with metadata about the sampling process. Sampling metadata can include the information that samples were collected under a “complete checklist” survey that reported all detected species on a checklist. Metadata can be used to build statistical models of distributions while accounting for sampling behaviors.

On the other hand, many participatory science observations are “unstructured”, meaning that observers do not report their sampling choices, and therefore effort and behaviors are unknown. Unstructured participatory science datasets are replete with unobserved, nonrandom variation in the data generating process (Di Cecco et al. 2021; Hughes et al. 2021). These data are useful for placing bounds on species ranges and for generating evidence of species behaviors and interactions from participant-generated photographs (Li et al. 2021). However, estimation of encounter rate or occupancy probability from these data is typically avoided (Johnston et al. 2022), precluding their use in large-scale monitoring of populations. The most prominent such unstructured participatory science dataset is iNaturalist (www.inaturalist.org), an app- and web-based platform for sharing biodiversity observations. For most taxonomic groups, including butterflies and many other invertebrates, far more data are generated from unstructured surveys than from semi-structured ones. Making use of unstructured data could greatly enhance monitoring efforts for these taxa.

One major source of sampling variation in unstructured participatory science data is taxonomic bias, which results when participatory scientists preferentially report certain species more than others (Ward 2014; Troudet et al. 2017). Taxonomic bias in reporting encompasses two types: logistical bias and preference bias. Logistical biases occur because some species are easier to identify and document than others. Such bias is prominent on platforms like iNaturalist that require evidence of species encounters. During a sampling event an iNaturalist participant might observe a number of interesting butterflies of various species, but if that participant is able to take a recognizable photograph of only one of those species, then only that single species observation will appear in iNaturalist. Drivers of logistical bias include body size (Barbato et al. 2021; Callaghan et al. 2021; Stoudt et al. 2022; Guedes et al. 2023), morphological distinctiveness (Koch et al. 2023; Mesaglio et al. 2023), and animal behavior (Le Roy et al. 2019). Likewise, preference biases
occur when observers choose to document certain detected species but not others out of personal interest (Troudet et al. 2017). Preference bias can reflect both individual and cultural values. Previous studies have proposed colorfulness and cultural values (such as avoidance of pest species) as drivers of preference bias (Stoudt et al. 2022).

As relatively distinct and well-liked invertebrates, butterflies are a unique taxonomic group in the participatory science data landscape (Ries and Oberhauser 2015). Studies of butterflies comprise some of the earliest and most impactful global change research (Parmesan et al. 1999; Eskildsen et al. 2013). In addition, butterflies are often one of the first points of contact between students and the natural sciences, as classroom rearing of butterflies in elementary schools is widespread. Butterfly-focused participatory science programs garner substantial public participation each year (Ries and Oberhauser 2015; Markee et al. 2021). For these reasons, butterflies are the best sampled insects in participatory science. Butterflies have both a dedicated semi-structured participatory science platform, eButterfly (www.e-butterfly.org; Prudic et al. 2017), which has documented more than 115,000 checklist sampling events, and the unstructured iNaturalist platform, with nearly 2.5 million detections. However, taxonomic biases in butterfly occurrence data are as yet unexplored.

When utilizing participatory science data, it is often useful to characterize how they compare to one another and to standardized datasets. Comparing data on human engagement with wildlife arising from different sampling protocols can shed light on the observation processes. These comparisons underpin much of the established literature regarding sampling biases (eg Troudet et al. 2017; Callaghan et al. 2021; Stoudt et al. 2022). Multi-dataset comparisons can highlight inaccuracies in large data aggregators like the Global Biodiversity Information Facility (GBIF) (eg Chollett and Robertson 2020), an increasingly important task as the use of these largely unvetted aggregators increases rapidly (Zizka et al. 2020). Quantifying differences between datasets of biodiversity observations is also a key step in integrating data from multiple sources within models (Pacifici et al. 2019).

Here, and to the best of our knowledge, we conducted the first study of species-level taxonomic bias in observations of any insect group by comparing two important programs collecting participatory science observations of butterflies. Using the generalized additive model (GAM) workflow presented by Stoudt et al. (2022), we explored whether each of 194 North American butterfly species is overreported or underreported in the unstructured, opportunistic iNaturalist dataset relative to the semi-structured eButterfly dataset. We then examined whether species traits and phylogenetic groupings, mediated by human preference and sampling logistics, explain which species are under- or overreported. On the basis of patterns observed in other taxa, we hypothesized that—in iNaturalist—colorful, large, and distinctly patterned species are overreported, while hard-to-identify species are underreported (see Appendix S1: Table S1). We present the first clear picture of trait- and taxonomy-associated bias in butterfly participatory science sampling and provide a species-specific quantified estimate of this bias that can be used to reduce bias in downstream analyses of iNaturalist data.

### Methods

#### Species detection rates

All iNaturalist “research-grade” observations of Lepidoptera were retrieved from the GBIF in April 2023 (GBIF 2023), while eButterfly records were obtained via direct request. We only considered eButterfly detections associated with complete checklist sampling events. Observations of butterflies during complete checklists reflect detection rates with no taxonomic bias, as all detected and identifiable species are reported. Although eButterfly observations can be uploaded without photographic evidence, 34% of observations are associated with photographs. In both datasets, we considered all observations of species in five families of North American butterflies (Nymphalidae, Lycaenidae, Papilionidae, Pieridae, Hesperiidae) occurring in the years 2000–2021 in the contiguous US and Canada, excluding the Canadian Yukon, Nunavut, and Northwest Territories due to low data density.

We divided the continent into two regions, West and East, to reflect differences in species composition between the two regions, and because the interior of the continent is largely unsampled in eButterfly (see Appendix S1: Figure S1). In each region, we aggregated both datasets to a spatial grid of regular hexagons with a long radius of 20 km and subdivided by year to explicitly accommodate temporal variation in detection rates (see Appendix S1: Panel S1). Because sampling density in eButterfly and iNaturalist followed different temporal patterns during the study period, matching reporting rates between the two datasets in both time and space was necessary to ensure that observed differences stemmed from differences in reporting rates and not from butterfly demography. In each cell-year, we counted the number of observations of each species. Each iNaturalist research-grade observation constituted a single species observation event. For equivalency, each species record on each eButterfly complete checklist counted as a single species observation event (Stoudt et al. 2022).

In each region, we modeled all species that appeared in at least 25 cell-years in each of the two datasets, a threshold of coverage necessary to estimate a spatiotemporally explicit model of overreporting: 149 species in the East region and 88 species in the West region met this criterion, including 43 species qualifying for modeling in both regions.

#### Trait data

For each butterfly species, we aggregated trait data representing hypothesized drivers of taxonomic bias. We obtained morphological traits representing the body size,
coloration, and wing features of each species. From the LepTraits dataset (Shirey et al. 2022), we retrieved each species’ wingspan in centimeters. We also retrieved the color and wing features of each species by parsing the descriptions of each species from the Butterflies and Moths of North America (BAMONA) website (www.butterfliesandmoths.org). We extracted all named colors (eg “red,” “yellow,” “black”) and features (eg “tail,” “spot,” “stripe”) from the identification field, then manually checked and corrected extracted values. Using these raw color and wing feature data, we calculated two derived trait metrics: color diversity (the number of colors present on the butterfly) and feature diversity (the number of features present on the butterfly) to capture potential associations between butterfly appearance and preference bias.

To represent the logistical constraint of identifying species from photographs, we calculated an index of “ease of identification.” From the iNaturalist website, we recorded, for each genus, the fraction of all observations that were research grade, giving the rate at which individuals in a genus can be identified to species, a proxy for recognizability (Mesaglio et al. 2023). To represent the effect of species rarity, we took the natural logarithm of the number of total observations submitted to eButterfly checklists for each species, which we termed “prevalence.” To investigate a potential effect of migratory status on overreporting, we retrieved species migratory status from a global review (Chowdhury et al. 2021). All species for which Chowdhury et al. (2021) found evidence of full or partial migration were considered migratory. The six main traits—prevalence, ease of identification, feature diversity, color diversity, wingspan, and migratory status—were not highly correlated with one another across species. All traits except species migratory status were centered and scaled for use in the meta-analysis.

Estimating species-level overreporting indexes

We estimated an overreporting index for each species in each region while accounting for variation in both space and time. First, we fit GAMs with quasi-binomial error distributions for each combination of species, dataset, and region admitted for analysis using the R package mgcv (v1.8.42; Wood 2011) (see Appendix S1: Panel S1 for full model definition). For each species–region–dataset combination, we obtained the spatiotemporally smooth logit-scale predicted reporting rate in each cell-year for each dataset. We then subtracted the predicted eButterfly reporting rate from the predicted iNaturalist reporting rate in each cell-year for each species–region pair. Finally, we took the median of these differences across space and time, yielding an “overreporting index” for each species–region. We used the parametric bootstrap approach described by Stoudt et al. (2022) to calculate 95% confidence intervals (CIs) for each overreporting index. We based inference across species on P values adjusted to control the false discovery rate at 0.05 (Benjamini and Hochberg 1995).

Summary analysis

We used a second-stage analysis to look for evidence that patterns in reporting bias across species were associated with taxonomic groupings and traits. Second-stage analyses were designed to model patterns in reporting bias across four dimensions—taxonomy, traits, colors, and features—while propagating uncertainty in species-level overreporting indices (Bürkner 2017; Stoudt et al. 2022) (see Appendix S1: Panel S1 for full model definitions). First, we fit a model estimating family-specific intercepts for each of the five taxonomic families represented in the species pool. Second, we fit a model estimating the effects of six species-level traits on overreporting: prevalence, ease of identification, feature diversity, color diversity, wingspan, and migratory status. Third, we estimated the effects of a butterfly’s wing having each of 11 colors (black, gray, white, brown, red, orange, yellow, green, blue, pink, and purple). Finally, we estimated the effects of a butterfly’s wing having each of five features: eyespots, bands/stripes, checkers, tails, or spots. All summary models were estimated using the R package brms (Bürkner 2017).

Results

We estimated an overreporting index with uncertainty for 149 and 88 species in the East and West regions, respectively. Across 194 unique species, 34 were associated with underreporting and 53 with overreporting. In the West region, we identified nine species that were significantly underreported in iNaturalist relative to eButterfly and 17 overreported species. In the East, we identified 31 significantly underreported and 39 overreported species (Figure 1). Of the 43 species studied in both the East and West regions, none gave contrasting inference between regions.

In two taxonomic families of butterflies, Lycaenidae (β 95% CI = 0.077, 0.302) and Nymphalidae (β 95% CI = 0.094, 0.260), species were typically overreported, whereas species in the family Pieridae (β 95% CI = –0.455, –0.174) were typically underreported. We did not find statistically clear evidence of over- or underreporting of species in the remaining two families, Papilionidae (β 95% CI = –0.006, 0.392) and Hesperiidae (β 95% CI = –0.155, 0.033). Model results matched a simple tally-based assessment of species patterns (Figure 2).

In a trait-based summary analysis, ease of identification (β 95% CI = 0.051, 0.164) and feature diversity (β 95% CI = 0.027, 0.136) were positively associated with species overreporting (Figure 3a). Four other traits—prevalence, color diversity, wingspan, and migratory status—were not credibly associated with overreporting.
We found that species with yellow (β 95% CI = –0.246, –0.015) and pink (β 95% CI = –0.656, –0.026) coloration were more underreported in a color-only model (Figure 3b), whereas a feature-only model suggested that species with spots (β 95% CI = 0.026, 0.278) and wing tails (β 95% CI = 0.004, 0.321) were more overreported (Figure 3c). Results are summarized in Appendix S1: Table S2.

**Conclusions**

Our analysis reveals widespread taxonomic bias in iNaturalist observations of butterflies. We identified 53 overreported and 34 underreported species of North American butterflies. These patterns closely followed taxonomic family delineations. In our analysis, families of the nymphalids and lycaenids tended to be overreported, whereas members of the family Pieridae tended to be underreported (Figures 1 and 2). We also identified strong trait associations explaining taxonomic bias. Ease of identification was the clearest logistical driver of reporting bias, whereas feature diversity was the clearest preference-based driver of reporting bias.

Species identification is a logistical hurdle when observing butterflies. Certain butterfly groups are especially challenging to distinguish in the field and from photographs, including members of the genera *Colias* (sulphurs) and *Erynnis* (duskywings). When identification depends on subtle features that may be obscured or unphotographed, indistinct species become underreported (Mesaglio et al. 2023). This effect is likely more important for datasets like iNaturalist, which require these features to be captured in photos, as opposed to platforms like eButterfly where only the observer is responsible for species identification. Logistical biases driven by the difficulty of species-level field identification are likely even more important in less visually distinct insect clades.

Participatory science programs seeking to ameliorate logistical bias in insect reporting could train local observers in best practices for field identification.

Individual users exhibit preferences for reporting species based on morphological traits. Our results indicate that feature diversity is strongly associated with preference bias in butterflies. Notably, color diversity was not a strong predictor of reporting preferences across butterfly species despite research showing colorfulness to be a driving factor in preferential reporting among birds (Stoudt et al. 2022). User preference and specialization may in part explain family-level differences in overreporting as well; individuals who contribute plant and
Bias in participatory science butterfly data

insect records to iNaturalist tend to be highly taxonomically specialized (Di Cecco et al. 2021), leading to greater iNaturalist reporting rates for some taxonomic families. Underreporting of Pieridae in particular may reflect cultural values, given that many species within this family are perceived to be garden pests. Logistical bias due to unmeasured behavioral traits, such as flight style or the vegetation strata of typical flight (e.g., high canopy flyers versus understory fliers), may also contribute to family-level patterns in overreporting.

![Phylogenetic tree of butterfly species](image)

**Figure 2.** A phylogenetic tree of butterfly species studied. Each point represents one species; blue points represent significantly overreported species and orange points represent significantly underreported species. Each slice of the circle corresponds to a taxonomic family (five in all), within which species are grouped; blue and orange represent whether the family mean was credibly different from zero in a family-level meta-analysis, whereas gray represents cases where the family mean was not credibly different from zero. Taxonomic families are also represented by the shape of each point. In addition to family-level patterns, we observed apparent clustering of under- and overreporting within families. For species with overreporting indices estimated in both regions, only the East region’s value is used. This phylogenetic tree is based on the phylogeny from Earl et al. (2021) (see Appendix S1: Panel S1).

Our modeling approach required species that met a minimum threshold of occurrence (at least 25 cell-years) to enable meaningful comparisons between datasets. As such, spatially restricted or rare species were pre-filtered from downstream analyses, and whether reporting trends hold for these species is unclear. Although a lack of phylogenetic signal for range size has been documented in a subset of North American butterflies (Lewthwaite et al. 2018), it is possible that this pre-filtering step disproportionately removed species from one taxonomic group versus another, or disproportionately...
removed species with certain traits. An additional challenge in estimating the effects of traits was the need to represent each species’ appearance with a single set of traits. Individual butterflies within a species may vary dramatically in morphology due to sexual dimorphism or metamorphosis. In our meta-analysis, we used traits based on typical adults (Shirey et al. 2022). Estimating sex- and life-stage-specific overreporting would require additional data, as eButterfly does not target caterpillars for sampling and does not collect data on sex. Similarly, most iNaturalist metadata do not indicate species life stage or sex.

Butterflies are a unique taxonomic group and an ideal model taxon for understanding how variation in human preference influences public engagement in the context of conservation. An understanding of aesthetic drivers of preference is critical for rigorous conservation; although aesthetics is not an efficient criterion for conservation prioritization, in practice, for butterflies and other species, human preferences drive the allocation of resources (Colléony et al. 2017; Lepczyk et al. 2023). This study provides a clear picture of species-level variation in public engagement with an important insect group.

Insect monitoring is entering a new era. Participatory data collection programs are generating unstructured observations at ever-increasing rates. Actual use of these data for monitoring populations and studying ecological systems across large scales requires a better understanding of their inherent systematic biases. Our findings suggest two important paths forward for ecologists seeking to use participatory science data to support butterfly conservation. First, models attempting to estimate distributions and abundance trends could be de-biased by including information on overreporting in the modeling framework. For example, the species-specific reporting indices we estimated could be used as an offset on observer bias in occupancy-detection models, which are increasingly fit to presence-only data (Shirey et al. 2023). Second, targeted local outreach could influence how iNaturalist data are generated in particular systems. Coordinators of participatory science efforts can boost engagement by highlighting engaging species, while simultaneously training participants to observe and identify underreported species that would otherwise go ignored. Given that recognizability is an important determinant of overreporting, coordinators could train participants in taking careful photos of necessary diagnostic features on hard-to-differentiate species. Combining these strategies for using participatory data to monitor butterflies and other insects could lead to a sea change in how insect species are monitored.

Acknowledgements

Thanks to eButterfly for providing the primary data; iNaturalist, Butterflies and Moths of North America (BAMONA), and the LepTraits team for producing data critical to this analysis; and to the countless eButterfly and iNaturalist participants, identifiers, and verifiers who generated the data used in this study, and whose passion for engaging with the natural world motivates a new era of biological data science. Thanks also to C Boettiger for providing computational support. BRG was supported by a US National Science Foundation (NSF) Graduate Research Fellowship (grant #1752814); VS was supported by an NSF Graduate
Research Fellowship (grant #1937959) and a David H Smith Conservation Research Fellowship; and LMG was supported by the Wrigley Innovation Fellowship and the Women in Science and Engineering (WiSE) program at the University of Southern California.

**Data Availability Statement**

All data and code used in this study have been made available. Code for reproducing all analyses, as well as modeled outputs and all input data except iNaturalist, can be found at [https://github.com/dochvam/butterfly-engagement-reproducible](https://github.com/dochvam/butterfly-engagement-reproducible). See the repository README for instructions on obtaining the publicly available iNaturalist data. At the request of eButterfly, global positioning system (GPS) coordinates for records of species that were considered sensitive (listed as Species At Risk in Canada or the US) were removed or obscured in the publicly available dataset we posted in accompaniment of this manuscript. Please contact the corresponding author with any questions regarding the code.

**References**


Supporting Information

Additional material can be found online at http://onlinelibrary.wiley.com/doi/10.1002/fee.2783/supplinfo

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.