Behavioral Ecology

The official journal of the
ISBE
International Society for Behavioral Ecology

Behavioral Ecology (2018), 00(00), 1–11. doi:10.1093/beheco/ary017

Original Article

Comparing colors using visual models

Rafael Maia[1] and Thomas E. White[2]

[1]Department of Ecology, Evolution and Environmental Biology, Columbia University, 1200 Amsterdam Avenue, New York, NY 10027, USA and [2]School of Life and Environmental Sciences, University of Sydney, Camperdown, Sydney, NSW 2006, Australia

Received 22 August 2017; revised 25 December 2017; editorial decision 6 January 2018; accepted 9 February 2018.

Color in nature presents a striking dimension of variation, though understanding its function and evolution largely depends on our ability to capture the perspective of relevant viewers. This goal has been radically advanced by the development and widespread adoption of color spaces, which allow for the viewer-subjective estimation of color appearance. Most studies of color in camouflage, aposematism, sexual selection, and other signaling contexts draw on these models, with the shared analytical objective of estimating how similar (or dissimilar) color samples are to a given viewer. We summarize popular approaches for estimating the separation of samples in color space and use a simulation-based approach to test their efficacy with common data structures. We show that these methods largely fail to estimate the separation of color samples by neglecting 1) the statistical distribution and within-group variation of the data and/or 2) the discriminability of groups relative to the observer's visual capabilities. Instead, we formalize the 2 questions that must be answered to establish both the statistical presence and theoretical magnitude of color differences, and propose a 2-step, permutation-based approach that achieves this goal. Unlike previous methods, our suggested approach accounts for the multidimensional nature of visual model data and is robust against common color-data features such as heterogeneity and outliers. We demonstrate the pitfalls of current methods and the flexibility of our suggested framework using an example from the literature, with recommendations for future inquiry.

Key words: **vision, dimorphism, polymorphism, mimicry, crypsis, multivariate statistics.**

INTRODUCTION

The study of color in nature has driven fundamental advances in ecology and evolutionary biology (Cuthill et al. 2017). Color is a subjective experience, however, so substantial effort has been dedicated to measuring and representing colors "objectively" (Garcia et al. 2014; Johnsen 2016) through visual models that consider the perspective of ecologically relevant viewers (Kemp et al. 2015; Renoult et al. 2017). These models have significantly advanced the study of color traits by allowing researchers to account for the factors influencing the generation and reception of visual information, such as the structure of signals and viewing backgrounds, the properties of veiling and incident light, and the attributes of visual systems (Chittka 1992; Vorobyev and Osorio 1998; Kelber et al. 2003; Endler and Mielke 2005).

Several forms of visual models are currently used, which vary in their assumptions about the nature of visual processing (Chittka 1992; Vorobyev and Osorio 1998; Endler and Mielke 2005). These models function by delimiting a color space informed by the number and sensitivity of photoreceptors in an animal's retina (Renoult et al. 2017). Individual colors are then represented in this space as points, with their location determined by the differential stimulation of the viewers' receptors.

This color space representation is convenient for several reasons. It offers an intuitive way of analyzing phenotypes that we cannot measure directly: we can estimate how animals with different visual systems "see" different colors by representing them in a Cartesian coordinate system, producing a receiver-dependent morphospace (Kelber et al. 2003; Renoult et al. 2017). Furthermore, it allows estimating how similar or dissimilar colors are to a given observer, by measuring the distance between color points in its color space (Vorobyev et al. 1998; Vorobyev and Osorio 1998; Endler and Mielke 2005). Crucially, we can test and refine these models using psychophysical data (e.g. Maier 1992; Vorobyev et al. 2001; Dyer and Neumeyer 2005; Garcia et al. 2017), to estimate the magnitude of color differences and ultimately predict whether an observer could effectively discriminate pairs of colors (Chittka 1992; Vorobyev and Osorio 1998). This final point is critical to many tests of ecological and evolutionary hypotheses, such as the efficacy of camouflage (Pessoa et al. 2014; Troscianko et al. 2016), the presence of polymorphism or dichromatism (Schultz and Fincke 2013; Whiting et al. 2015), the accuracy of mimicry (O'Hanlon et al. 2014; White et al. 2017), the extent of signal variability among populations or species (Delhey and Peters 2008; Rheindt et al. 2014; Dalrymple et al. 2015), or the effect of experimental

Address correspondence to R. Maia. E-mail: rm3368@columbia.edu

manipulations (Barry et al. 2015; White and Kemp 2017). At the heart of these inquiries lies the same question: how different are these colors to the animal viewing them?

Challenges in estimating the discriminability of color samples

The receptor noise-limited model of Vorobyev and Osorio (1998) has proven particularly useful for addressing questions of discriminability and color difference. The model is focused on receptor-level processes, and assumes that chromatic and achromatic channels operate independently (which does not necessarily hold beyond the receptor level in some species, such as humans; Nathans, 1999), that color is coded by $n - 1$ unspecified opponent mechanisms (where n is the number of receptor channels), and that the limits to color discrimination are set by noise arising in receptors (Vorobyev and Osorio 1998; Vorobyev et al. 1998). This noise is dependent on the receptor type and abundance on the retina which, along with Weber's law $\left(k = \frac{\Delta I}{I}\right)$ more generally, ultimately establishes the unit of Just Noticeable Differences (JND; Vorobyev et al. 2001). Distances calculated in this manner correspond to the Mahalanobis Distance D_M, and represent distances between points standardized by the Weber fraction; i.e. $\frac{signal}{noise}$ (Clark et al. 2017).

It follows that values lower than 1 JND $\left(\frac{signal}{noise} < 1\right)$ are predicted to be indistinguishable, while values greatly above this threshold are likely distinct. This provides a useful standard for estimating the similarity of groups of points in color space: the greater the distance between colors, the less alike they are. If differences are, on average, above an established threshold, then we can consider the groups different: sexes dichromatic, mimetism imperfect, and crypsis ineffective. This offers a clear link between variation and classification within a sensory framework and has been widely used for this purpose (Delhey and Peters 2008; Schultz and Fincke 2013; O'Hanlon et al. 2014; Barry et al. 2015; White and Kemp 2017; White et al. 2017).

To adequately compare samples of colors, however, it is necessary to determine if the average distance between them is both statistically and biologically meaningful (i.e. above-threshold; Endler and Mielke 2005). Commonly, an "average colour" for each group is derived by taking a mean reflectance spectrum or by averaging their position in color space. In either case, the color distance between groups is then calculated from these mean quantum catches per-receptor per-group—their centroids in multivariate space (Figure 1, bold arrow). However, the centroid obtained from arithmetic means of receptor coordinates is not an appropriate measure of location for this purpose, since color distances are perceived in a ratio scale (Cardoso and Gomes 2015). Instead, the geometric mean must be used. Furthermore, since the result is a single value representing the multivariate distance between-group means, there is no associated measure of uncertainty or precision that would allow for the statistical testing of differences between samples (e.g. Avilés et al. 2011; Burns and Shultz 2012; Maia et al. 2016).

An alternative approach calculates the pairwise distances between all points in group A and group B, then averages these distances to obtain a mean distance between groups (Figure 1, thin arrows; e.g. Dearborn et al. 2012; Barry et al. 2015). In cluster analyses, this is called the "average linkage" between groups (Hair et al. 1998). This is an appealing method, providing measures of variation among distances, and thus a *t*-test or equivalent can be used to test if differences are greater than a given threshold. The average linkage, however, is also inadequate because it conflates within- and among-group variation. This is because Euclidean distances (and by extension JND's) are translation-invariant: they ignore the position of points in color space and the direction of the distance vectors, reflecting only the magnitude of differences between 2 points. Therefore, the average linkage reduces to a measure of spread, and will scale with both within- and between-group distances (Figure 1, inset).

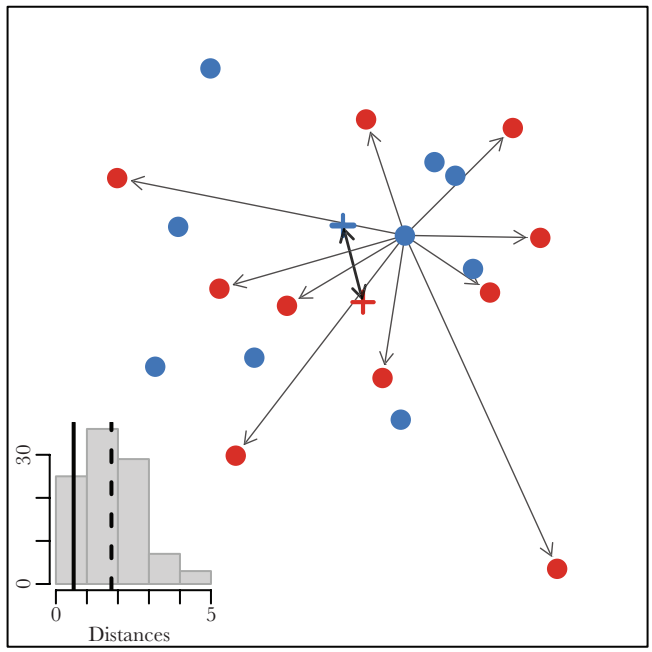

Figure 1
The link distance (ie, average pairwise distance between groups) conflates within- and among-group variation. Here, 2 samples were drawn from the same simulated distribution. Thin arrows represent distances between a random point in the first sample (blue) and all points from the second sample (red), all of which are greater than the distance between the geometric means of the 2 samples ("x", bold arrows). Inset shows the histogram of pairwise distances among groups, and how their average (dashed line) is greater than the mean distance (bold line).

As these issues show, hypotheses of discriminability and color difference have primarily focused on testing whether the difference between samples is above a theoretical threshold. However, the convenience of such thresholds belies the fact that simply comparing means between groups is not sufficient to infer, statistically, whether samples are different. To answer if 2 groups are different, one must compare the variation between- and within-groups. This is particularly problematic in the case of colors that function as signals in social interactions (e.g. Kemp and Rutowski 2011). For a trait to function in this context, the observer must be able to tell signals of "low" and "high" quality apart. This means that, by definition, *most pairs of individuals should be readily distinguishable*. The trait must be highly variable and color distances should be above the threshold of discrimination (Delhey et al. 2017), otherwise no information can be extracted by an observer comparing phenotypes.

Consider a hypothetical species that uses color in mate choice but is not sexually dichromatic (Figure 1). In this species, color is highly variable and, on average, pairs of individuals are

discriminable, but there is no *consistent* male–female difference. Therefore, if a researcher sampled this species and calculated the average distance between all pairs of individuals, regardless of sex, these should be largely greater than 1 JND. However, if they took separate samples of males and females, then all pairwise distances (the average linkage) between sexes will be also greater than 1 JND, despite them being sampled from the same (statistical) population.

The limitations of current methods for comparing color space distributions

Several methods have been proposed to avoid the aforementioned issues by accounting for the relative distributions of samples in color space. Eaton (2005), for example, noted that within-group variation influenced the conclusions on the extent of avian dichromatism, and thus tested for intersexual differences in photon catches separately for each receptor. However, this ignores the multivariate nature of visual model data by failing to account for multiple comparisons and correlations among receptor catches (which are critical, since any *n*-receptor visual system can be represented in $n - 1$ dimensions; Kelber et al. 2003).

An alternative, multivariate metric suggested by Stoddard and Prum (2008) is the volume overlap. In this approach, the volume occupied by a sample of colors is estimated from its enveloping convex hull, and separation between samples is inferred from their overlap. Stoddard and Stevens (2011) used this metric to show that a greater overlap in color volume between cuckoo and host eggs is associated with lower rejection of parasitic eggs. This approach is appealing because it considers the distribution of color points in multivariate space, though there are limits to its interpretation: 1) there is a lower bound to group separation (i.e. if samples do not overlap, there is no distinction between cases where samples are near or far apart in color space) and 2) it is unclear how variation in volume overlap should be interpreted biologically (e.g. how biologically relevant is the difference between 20% or 40% overlap?). It is also particularly sensitive to outliers, because the volume defined by a convex hull does not lend itself to a probabilistic interpretation, leading to the often-unacknowledged assumption that the sampled data reflects the true boundaries of the population (however, "loose wrap" hypervolumetric methods exist; to our knowledge, these have not been applied to color studies; Blonder et al. 2017). Finally, in its original implementation this method does not consider receptor noise or discrimination thresholds (but incorporating this is straightforward; see below).

The most robust attempt at comparing distributions of colors was proposed by Endler and Mielke (2005), who devised a nonparametric rank distance-based approach based on the least sum of Euclidean distances, compared through multi-response permutation procedures (LSED-MRPP). This multivariate approach is powerful because it calculates an effect size based on the relationship of between- and within-group distances. However, this single statistic captures differences between samples not only in their means, but also in their dispersion and correlation structure (i.e. shape; Endler and Mielke 2005). Like other distance-based methods, it is sensitive to confounding variance heterogeneity among samples when testing for differences in *location* (Warton et al. 2012; Anderson and Walsh 2013). Despite its considerable strengths, this method has seen little adoption over the last decade, largely due to limitations in implementation and accessibility.

The shortcomings of these methods reflect the fundamental fact that the question of discriminability actually represents a test of 2 hypotheses that are seldom formally distinguished: 1) that the focal samples are statistically distinct and 2) that the magnitude of their difference is greater than a psychophysical threshold of detection. Most approaches will test one, but not both, of these hypotheses through their respective nulls, and often with no estimate of uncertainty. We explore these issues using a simulation-based approach by testing the efficacy of popular methods in detecting the separation of groups in color space. We then propose a flexible solution that avoids these problems, demonstrating its utility using an example from the literature.

METHODS

Simulation procedures

To compare methods for detecting group separation in color space, we simulated data analogous to that obtained from applying an avian visual model to spectral reflectance data. Birds are tetrachromatic (Hart 2001), and colors will thus be represented by the quantum catches of its 4 photoreceptors (though the procedure followed here can be applied to visual systems with any number of receptors). For each replicate, we simulated 2 samples defined by 4 variables (*USML* photoreceptors) taken from log-normal distributions (since quantum catches are non-negative and noise-corrected distances follow a ratio scale, as defined by the Weber fraction, described above). We generated samples following 2 different scenarios: first, we simulated varying degrees of separation (i.e. effect sizes) to evaluate the power and Type I error rates of the approaches tested. Second, we simulated threshold conditions to evaluate the performance of different approaches in correctly classifying whether samples are above-threshold.

For the first set of simulations (power and error rates), we simulated the quantal catch of each photoreceptor i for the first sample (group A) by drawing from a log-normal distribution with mean μ_{iA} seeded from a uniform distribution $U(0,10)$, and standard deviation proportional to the mean: $\sigma_i = a_i \mu_{iA}$, with $a_i \sim U(0,0.5)$ (note that, for these simulations, μ and σ refer to the mean and standard deviation of the random variable itself, not in log scale). To generate 2 samples with varying degrees of separation proportional to the within-group variance, we used a multivariate effect size S obtained by calculating a constant $d_i = \frac{S}{\sqrt{n}} \bar{\sigma}_i$, where n is the number of photoreceptors (in this case, 4 and $\bar{\sigma}_i$ is the standard deviation of the sample. We then drew a second sample (group B) defined by $\mu_{iB} = \mu_{iA} + d_i$ and σ_i. Thus, our simulations effectively produced 2 samples with Mahalanobis Distance $D_M \sim S$ (calculated as the distance between centroids of the 2 groups weighted by their pooled variance-covariance matrix). We simulated data for $S = \{0, 0.1, 0.25, 0.5, 0.75, 1.0, 1.5, 2, 2.5, 3.0\}$ (Figure 2), replicated 200 times for sample sizes $N = \{10, 20, 50, 100\}$ each.

For the second set of simulations (threshold conditions across a range of within-sample variation), we followed a similar procedure. Group A was sampled from a log-normal distribution with $\mu_{iA} \sim U(0,10)$ while σ_i was taken from an exponential distribution $\sigma_i \sim Exp(\lambda = 1)$. To obtain a second sample, group B, that was separated from group A with an average approximate distance of 1 JND given a Weber fraction of 0.1 for the long-wavelength photoreceptor (Vorobyev et al. 1998), we drew from log-normal distributions with $\mu_{iB} = d_i \mu_{iA}$, where $d_i \sim U(0.88,1.12)$, resulting in an average distance between geometric means (hereafter, "mean distance") of 1.11 (95% quantiles: 0.35–2.77 JND) and within-group

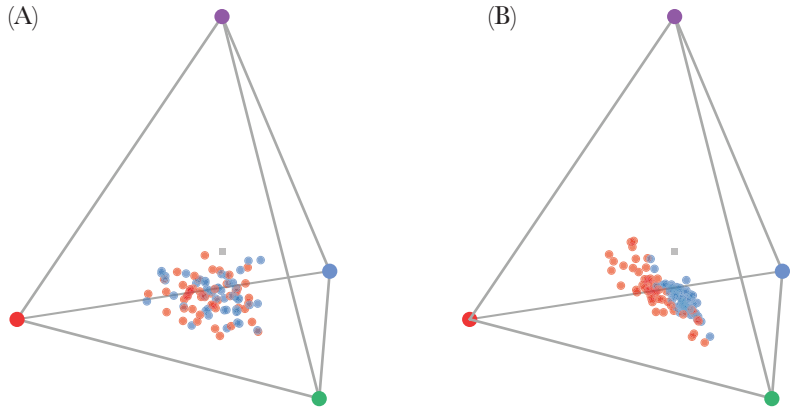

Figure 2
Example simulated data for the 2 groups (red, blue) in a tetrahedral color space. Shown here are data with sample size $N = 50$ and effect size (A) S = 0 and (B) S = 3.

average pairwise distance of 4.46 (95% quantiles: 1.03–11.10 JND) after 1000 replicates.

After the 2 groups were simulated, we used the R package pavo (Maia et al. 2013) to calculate colour distances using relative receptor densities of $\{U, S, M, L\} = \{1, 2, 2, 4\}$ and Weber fraction for $L = 0.1$. We calculated the within-group average pairwise distance, as well as the distance between sample geometric means.

We then used four procedures to test for differences between groups. First, we used a distance-based PERMANOVA (hereafter "distance PERMANOVA") using the *adonis* function from the R package vegan (Oksanen et al. 2007). This non-parametric approach uses distances to calculate a pseudo-F statistic, simulating a null distribution by randomizing distances between observations (Anderson 2005). We recorded if the analysis was significant ($\alpha = 0.05$) using 999 permutations for the null, as well as the R^2 as an effect size estimate. Second, we obtained XYZ Cartesian coordinates based on "perceptually-scaled" (i.e. noise-corrected) distances (Pike 2012; functionally and mathematically equivalent to the receptor-noise limited space of Hempel de Ibarra et al. 2001) and applied a MANOVA test on these coordinates (hereafter "Cartesian MANOVA"). For simplicity, we used a sum of squares and cross-products matrix approach and calculated Pillai's trace and its associated P value (but see discussion and Electronic Supplementary Material for extensions of this approach). Third, we calculated the volume overlap between the 2 samples (relative to their combined volumes) in a tetrahedral color space defined by the receptors' relative quantum catches (thus disregarding receptor noise; Stoddard and Prum 2008). Finally, we calculated the volume overlap for the XYZ Cartesian coordinates based on noise-corrected distances, generating a color volume overlap that accounts for receptor noise.

SIMULATION RESULTS

Power and error rates

Both the distance PERMANOVA and the Cartesian MANOVA showed appropriate Type-I error rates, with about 5% of our simulations producing significant results when $S = 0$, even for small sample sizes (Figure 3). As expected, the power to detect small effects steadily increased as a function of sample size, with the distance PERMANOVA being overall more conservative than the Cartesian MANOVA (Figures 3 and 4).

The 2 approaches showed some disagreement, with between 10% and 15% of the simulations significant only in one of the 2 tests (Figure 4). This disagreement was not random, with the Cartesian MANOVA being more likely to be significant when the distance PERMANOVA was not than vice-versa (Figure 4A), at an approximately constant rate across sample sizes, and disagreement being concentrated at smaller effect sizes with increasing sample sizes (Figure 4B).

Focusing on $N = 50$ simulations, our results show that mean distance was positively associated with the effect size, and the threshold of significance using the distance PERMANOVA fell approximately at the 1 JND mark (Figure 5A; equivalent results are observed with the Cartesian MANOVA, not shown). Still, even around that threshold, significance is variable, showing that large within-group variation can lead to non-significant differences between groups despite among-group distances being above the theoretical perceptual threshold. Volume overlap also showed a (negative) association with effect size, but no specific threshold for significance is observed (e.g. both significant and non-significant results are observed for 20−60% overlap; Figure 5B).

Threshold scenarios

Since results from the distance PERMANOVA and the Cartesian MANOVA were comparable, we focus on the former due to the convenience of the R^2 statistic describing among-group separation (but see Discussion for comments on the use of these approaches). Simulations produced a wide range of outcomes, with non-significant and significant tests both above and below the theoretical threshold of 1 JND (Figure 6). In contrast with the power simulations above (Figure 5), the significance threshold did not match the theoretical perceptual threshold. As in the hypothetical example from the Introduction, 20.2% of the simulated cases were statistically indistinguishable despite having mean above-threshold distances (Figure 6, dark red). Likewise, 15.1% of the simulations produced samples that were statistically different, but where this difference was below threshold and was therefore likely undetectable to its observer (Figure 6, dark blue points). These results highlight the importance of considering both statistical separation and theoretical perceptual thresholds when testing the hypothesis that samples are discriminable.

Figure 6A shows that, intuitively, tests were significant when within-group differences were small relative to among-group

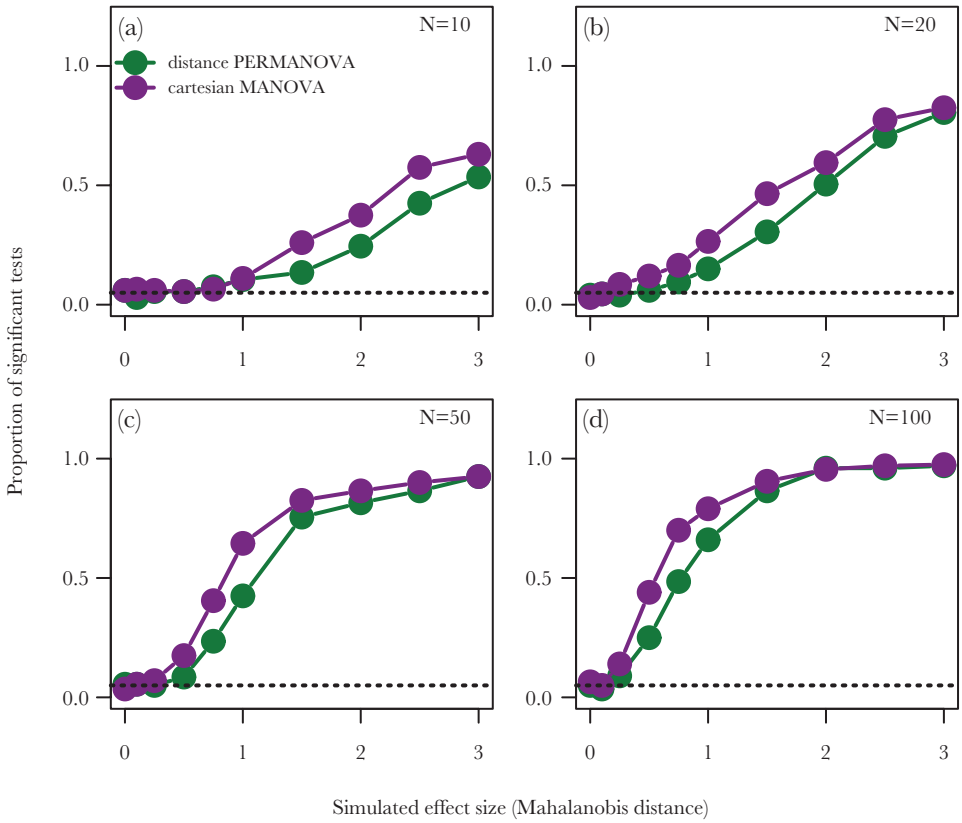

Figure 3
Power and Type I error rate of the distance PERMANOVA (green) and Cartesian MANOVA (purple). Panels show the proportion of simulations yielding significant results for each approach under different sample and effect sizes.

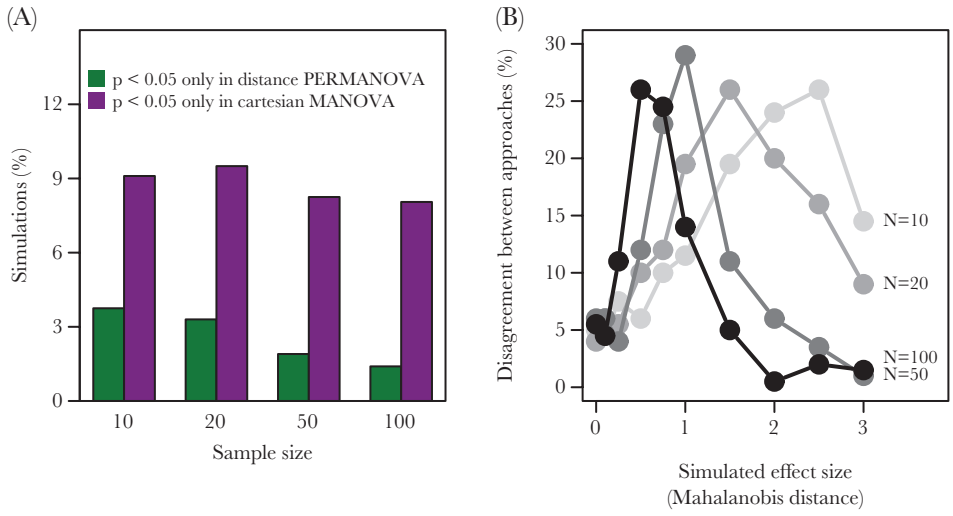

Figure 4
The disagreement between multivariate statistical approaches when testing for separation between samples in color space in relation to sample size (A) and effect size (B).

differences. However, nearly all simulations—including most significant results—fell below the 1:1 line when using the average link distance (i.e. the average pairwise distance) to describe intragroup variation. Significant results are obtained when the mean difference is up to 0.5 JND smaller than the within-group average link distance (Figure 6A, gray line intercept). Similarly, we can see that significant results can be obtained for fairly low levels of among-group separation, with R^2 as small as 3 or 4% (Figure 6B, horizontal line at 3%).

Though there is a negative association between R^2 and volume overlap (Figure 6C), the results show low overall consistency between approaches: for any given value of volume overlap, all possible outcomes of significance/threshold occur—even when the overlap between samples is zero (Figure 6C). In other words, even complete separation in color volumes can result in non-significant, below-threshold cases, since samples can be contiguous without overlapping in noise-corrected color space. Likewise, samples can

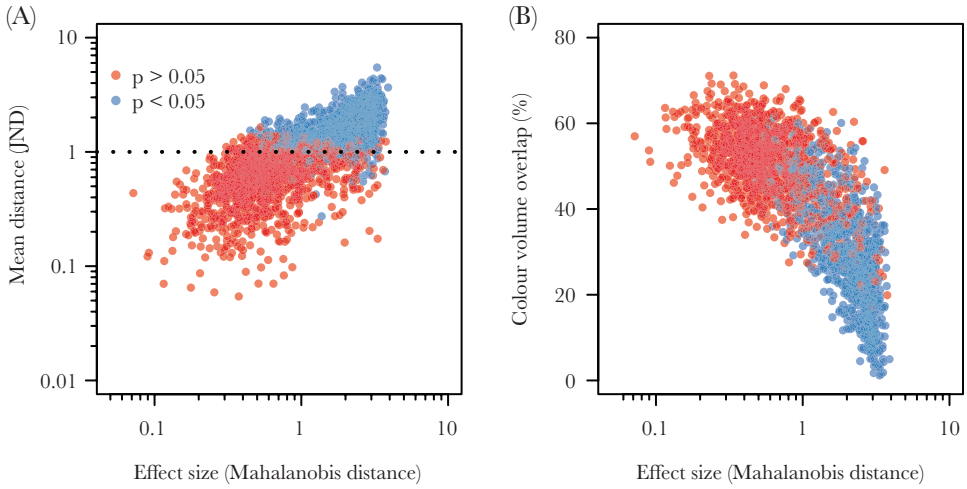

Figure 5
The association between effect size and (A) mean distance and (B) color volume overlap. Significant distance PERMANOVA results are in blue, whereas nonsignificant results are in red. Dotted line indicates the threshold of 1 JND.

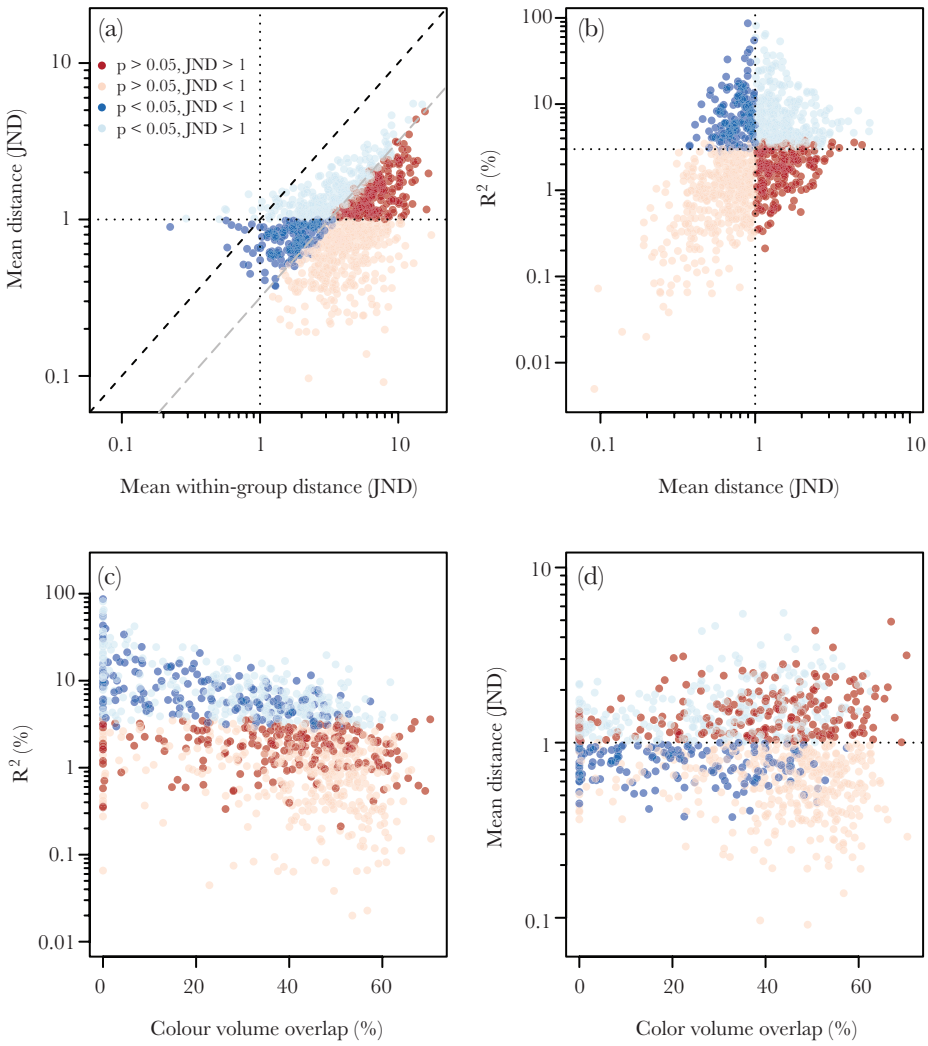

Figure 6
Results from threshold simulation. Red and blue denote nonsignificant and significant PERMANOVA tests, respectively, and light colors denote when that approach would yield the same inference as comparing mean distances to a threshold of 1 JND. Thus, dark blue points indicate a significant statistical test that does not reach the threshold of discriminability of 1 JND, whereas dark red points indicate a nonsignificant statistical test that nonetheless has a mean distance greater than 1 JND.

have high overlap but be entirely distinguishable statistically and perceptually. Further, there is no association between volume overlap and mean distance between groups (Figure 6D). These results were unaltered by considering receptor noise in the volume overlap calculation, since these are still strongly and positively correlated with their non-noise-corrected counterparts (see Electronic Supplementary Material).

A 2-step approach to estimate statistical and perceptual separation

As described previously, questions of discriminability and color difference require testing 2 distinct hypotheses: if samples are 1) statistically and 2) "perceptually" distinct. We, therefore, propose a 2-step answer to such questions, which explicitly formalizes these hypotheses. For the first question—are the samples *statistically separate* in color space?—we show that both a PERMANOVA using noise-corrected color distances (Anderson 2005; Cornuault et al. 2015), and a MANOVA using noise-calibrated Cartesian coordinates (Hempel de Ibarra et al. 2001; Pike 2012; Delhey et al. 2015) are well suited. Both exclude achromatic variation and properly account for the multivariate nature of the data. There is also minimal discrepancy between the two (Figures 3 and 4), so the decision between them may be informed by convenience and the structure of the data at hand.

Once the separation of samples is established statistically, a second question must be answered: is this separation predicted to be *perceptually discriminable*? The statistics calculated above cannot answer this, since effect sizes account for both among- and within-group variance. We, therefore, suggest this be tested independently, by estimating the distance in color space between-group geometric means rather than through the average pairwise distance or volume-overlap based metrics, which fail to accurately estimate group separation (Figures 1 and 6). One limitation to this statistic is the lack of any measure of uncertainty. To circumvent that, we suggest a bootstrap procedure in which new samples are produced through resampling (with replacement) of individuals of each group, from which geometric means and their distance are calculated. Repeating this procedure generates a distribution of mean distances, from which a confidence interval can be estimated. If the groups being compared are statistically different and this bootstrapped confidence interval does not include the theoretical threshold of adequate biological significance, one can conclude that the samples being compared are distinct and likely discriminable.

EMPIRICAL EXAMPLE: SEXUAL DICHROMATISM IN THE LEAF-NOSED LIZARD *CERATOPHORA TENNENTII*

Visually signaling animals often use distinct body parts for different purposes, such as social signaling to mates or warning predators (Johnstone 1995; Grether et al. 2004; Barry et al. 2015). The nature of intraspecific variation in color can thus inform their putative function, since selection may act differentially on signals used in different contexts. For example, traits subject to strong sexual selection in one of the sexes are often dimorphic, with one sex (typically males) expressing a conspicuous color pattern that is reduced or absent in the other (Kemp and Rutowski 2011; Bell and Zamudio 2012).

Dragon lizards (Agamidae) are known for variable coloration used in both social and anti-predator contexts (Somaweera and Somaweera 2009; Johnston et al. 2013). The leaf-nosed lizard *Ceratophora tennentii* has multiple discrete color patches, with apparent sex differences between body parts (Figure 7). Here, we draw on the data of Whiting et al. (2015), who recorded the spectral reflectance of 29 male and 27 female *C. tennentii* from four body regions (throat, labials, mouth-roof, and tongue). We used a tetrachromatic model of agamid vision to test for dichromatism among body regions from the perspective of conspecifics.

Following standard calculations for the log-linear receptor-noise model, we used the spectral sensitivity of *Ctenophorus ornatus* (λ_{max} = 360, 440, 493, 571 nm) as modeled according to a vitamin A1 template (Govardovskii et al. 2000; Barbour et al. 2002). We assumed a relative photoreceptor abundance of 1:1:3.5:6, and a coefficient of variation of noise yielding a Weber fraction of 0.1 for the long-wavelength cone (Loew et al. 2002; Fleishman et al. 2011). We tested each body region separately using PERMANOVA. As above, we used the R package pavo for visual modeling, and the adonis function in the R package vegan for PERMANOVAs.

We found a statistical difference between male and female throats (PERMANOVA: $F_{1,58} = 14.84$, $P < 0.01$) and labials ($F_{1,57} = 13.96$, $P < 0.01$; Figure 7A and B), but not for tongues ($F_{1,58} = 1.63$, $P = 0.22$) or mouth-roofs ($F_{1,55} = 0.52$, $P = 0.50$; Figure 7C and D). However, bootstraps of group separation suggest that intersexual differences in labial color are likely imperceptible to conspecifics (Figure 7E; though like all such predictions this requires behavioral validation). Our results therefore suggest the absence of dichromatism in all but throat color from the lizard perspective, despite statistical significance for the labial region. These results thus do not implicate sexual selection as a strong driver of intersexual color differences in these few body regions of *C. ornatus*.

DISCUSSION

Visual models offer a useful tool for quantifying the subjective perception of color, which—as the ultimate canvas for color-signal evolution—can offer valuable insight into a breadth of biological phenomena. It is therefore essential that statistical considerations of biological hypotheses take into account both natural variation in the compared samples as well as the limits to observer perception (as ultimately informed by behavioral and physiological data; Kemp et al. 2015). Here, we show that most methods typically fail to consider these aspects and propose a flexible alternative that explicitly addresses both.

The use of models that do not explicitly consider discriminability, such as the volume-overlap and segment-based analyses, is often justified on the basis of simplifying and relaxing assumptions about color perception, since intricate empirical work is required to estimate model parameters (Vorobyev and Osorio 1998; Olsson et al. 2015; Kelber et al. 2017). However, we contend that, on the contrary, some of these "simpler" methods actually make very strong latent assumptions, which are not supported by the empirical evidence. This includes the assumption that all cones contribute equally to color perception, that color discrimination is unequivocal (i.e. the magnitude of color difference does not affect discriminability) and that color differences follow an interval scale (as opposed to a ratio scale). Thus, we suggest that considering detectability relative to a threshold is essential for tests of discriminability. We emphasize, however, that this does not necessitate the use of the receptor-noise model specifically. Although we have focused on this popular approach here, particularly due to its utility for non-model organisms, a breadth of available modeling tools are capable of offering similar, and in some cases superior, insight (Kemp et al.

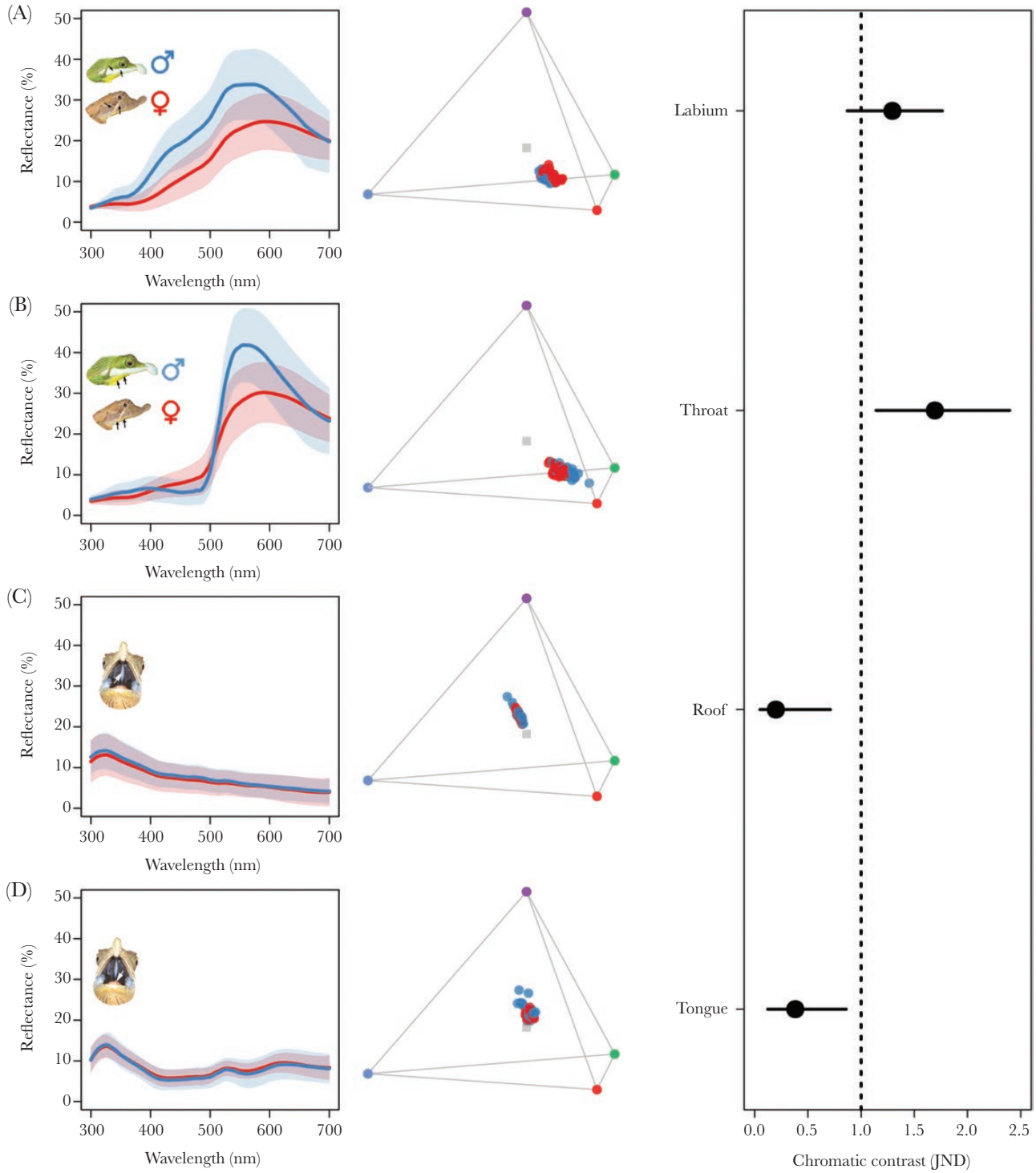

Figure 7
The mean (±SD) spectral reflectance of female (red) and male (black) (A) labial, (B) throat, (C) mouth-roof, and (D) tongue (left panels), and their color space distribution according in a tetrachromatic model of agamid vision (middle panels). Inset images indicate approximate sampling regions. The bootstrapped 95% CI's for mean distances between groups in color space (right panels). Partly reproduced, with permission, from Whiting et al. 2015.

2015; Price and Fialko 2018; Renoult et al. 2017). The hexagon model of Chittka (1992), for example, has been extensively tested and validated in honeybees, and may outperform the receptor-noise model when suitably parameterized (Garcia et al. 2017). It too offers a psychophysiologically informed measure of perceptual distance, as well as discrimination thresholds, and so may be readily applied within our suggested framework. Indeed, the 2-step approach we propose can be easily and directly extended to all such models.

Our simulations show that both the distance PERMANOVA and Cartesian MANOVA perform similarly well in statistically differentiating colors in perceptual space (Figure 3). Studies have pointed out

that distance-based methods perform poorly when the experimental design is unbalanced or when there is heteroscedasticity (though, among distance-based methods, PERMANOVA outperforms other approaches; Warton et al. 2012; Anderson and Walsh 2013). It is important to note that these are often common features of, and are applicable to, color data (Endler and Mielke 2005) and that these assumptions should be considered and verified. However, this might still be the most robust option for high-dimensional visual systems (e.g. Arikawa et al. 1987; Cronin and Marshall 1989), by reducing data to a single metric of distance. Recently, Delhey et al. (2015) advocated a similar MANOVA-based approach, by applying a Principal Component Analysis (PCA) to the noise-corrected Cartesian coordinates prior to the test. However, if all the principal components are used in the MANOVA, results will be numerically identical to directly using the XYZ coordinates (which is preferable, since it is often tempting to discard PC axes of low variance, which could be problematic given that those axes may be involved in group differentiation). Though we have focused on tests of differences in the "location" of colors in color space, we recognize that other characteristics—such as differences in dispersion and correlation structure, or the direction of variation among groups—might themselves be of biological interest, for which a PCA approach may be particularly useful.

The MANOVA approach can also be extended to multivariate generalizations of generalized linear models by using the noise-corrected Cartesian coordinates as response variables (Hadfield 2010). These models can relax the assumptions of heteroscedasticity by estimating the variance-covariance of the groups (Hadfield 2010) and can be extended to include various error and model structures, such as hierarchical and phylogenetic models (Hadfield and Nakagawa 2010). Still, these approaches will only test for the statistical separation in color space, so estimating the magnitude of that separation is still necessary. The bootstrapped distance provides an easy to interpret measure of uncertainty to the mean distance estimate. Under a Bayesian approach, the mean distance bootstrap can be substituted by estimating credible intervals for the distance between perceptually corrected Cartesian centroids from the posterior distribution, though this will be influenced by the priors adopted (Hadfield 2010, see Electronic Supplementary Material for an example).

Irrespective of the method used, it is essential to parameterize the underlying visual model appropriately (Garcia et al. 2017; Olsson et al. 2018). The Weber fraction and receptor densities chosen will strongly affect noise-corrected distances since they directly scale with the JND unit (Bitton et al. 2017). Further, even under adequate values of the Weber fraction, it is important to realize that the unit JND usually reflects psychophysical performance under extremely controlled conditions (Kelber et al. 2003; Olsson et al. 2015) and that more conservative estimates of 2–4+ JND may be more appropriate for ecological and evolutionary questions (Osorio et al. 2004; Martin Schaefer et al. 2007). Sensitivity analyses are also useful for exploring the robustness of conclusions against parameter variation, particularly in the case of non-model systems where such values are often assumed or drawn from related species (Bitton et al. 2017; Olsson et al. 2018). More broadly, we affirm recent (and ongoing) calls for pragmatism when drawing inferences from any such model (Marshall 2018; Olsson et al. 2018; Vasas et al. 2018). Color spaces are valuable tools, but ultimately demand ongoing feedback from physiological and behavioral tests to improve our understanding of complex biological phenomena.

Our results show that insight into the biology of color and its role in communication is best achieved by disentangling the implicit assumptions in questions of discriminability. By bringing these assumptions to light, our 2-step approach offers a flexible procedure for examining the statistical presence and theoretical magnitude of differences between color samples. We expect that it will bring exciting new perspectives on the role of color in intra- and inter-specific interactions and provide an efficient analytical framework for the study of color in nature.

IMPLEMENTATION AND DATA ACCESSIBILITY

Analyses and simulations can be found in github.com/rmaia/msdichromatism (publication version archived, doi: 10.5281/zenodo.1149585), and the described methods are fully implemented in the R package pavo as of version 1.3.1, available via CRAN. Key functions include "bootcoldist," which calculates the bootstrapped confidence intervals for mean distances, whereas "jnd2xyz" converts chromatic distances in JNDs to noise-corrected Cartesian coordinates. Multi-dimensional plotting options for noise-converted coordinates are also available. Analyses reported in this article can be reproduced using the archived data of Whiting et al. (2015).

SUPPLEMENTARY MATERIAL

Supplementary data are available at *Behavioral Ecology* online.

ACKNOWLEDGMENTS

We would like to thank Ruchira Somaweera for leaf-nosed lizard photographs, and Dan Noble, John Endler, and 2 anonymous reviewers for valuable discussion and comments on earlier drafts of the manuscript. This work was partially supported by a Junior Fellow award from the Simons Foundation to R.M., and an Australian Research Council grant (DP140140107) to T.E.W.

AUTHOR CONTRIBUTIONS

R.M. and T.E.W. conceived the ideas, designed methodology, analyzed the data, and wrote the manuscript.

Handling editor: Bob Wong

REFERENCES

Anderson MJ. 2005. Permutational multivariate analysis of variance. Auckland (New Zealand): Department of Statistics, University of Auckland. 26:32–46.

Anderson MJ, Walsh DC. 2013. Permanova, anosim, and the mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? Ecol Mono. 83:557–574.

Arikawa K, Inokuma K, Eguchi E. 1987. Pentachromatic visual system in a butterfly. Naturwissenschaften. 74:297–298.

Avilés JM, Soler JJ, Hart NS. 2011. Sexual selection based on egg colour: physiological models and egg discrimination experiments in a cavity-nesting bird. Behav Ecol Sociobiol. 65:1721–1730.

Barbour HR, Archer MA, Hart NS, Thomas N, Dunlop SA, Beazley LD, Shand J. 2002. Retinal characteristics of the ornate dragon lizard, *Ctenophorus ornatus*. J Comp Neurol. 450:334–344.

Barry KL, White TE, Rathnayake DN, Fabricant SA, Herberstein ME. 2015. Sexual signals for the colour-blind: cryptic female mantids signal quality through brightness. Fun Ecol. 29:531–539.

Bell RC, Zamudio KR. 2012. Sexual dichromatism in frogs: natural selection, sexual selection and unexpected diversity. Proc Biol Sci. 279:4687–4693.

Bitton PP, Janisse K, Doucet SM. 2017. Assessing sexual dicromatism: the importance of proper parameterization in tetrachromatic visual models. PLoS One. 12:e0169810.

Blonder B, Babich Morrow C, Maitner B, Harris DJ, Lamanna C, Violle C, Enquist BJ, Kerkhoff AJ. 2017. New approaches for delineating n-dimensional hypervolumes. Method Ecol Evol. doi: 10.1111/2041-210X.12865.

Burns KJ, Shultz AJ. 2012. Widespread cryptic dichromatism and ultraviolet reflectance in the largest radiation of neotropical songbirds: implications of accounting for avian vision in the study of plumage evolution. The Auk. 129:211–221.

Cardoso GC, Gomes ACR. 2015. Using reflectance ratios to study animal coloration. Evol Biol. 42:387–394.

Chittka L. 1992. The colour hexagon: a chromaticity diagram based on photoreceptor excitations as a generalized representation of colour opponency. J Comp Phys A. 170:533–543.

Clark RC, Santer RD, Brebner JS. 2017. A generalized equation for the calculation of receptor noise limited colour distances inn-chromatic visual systems. R Soc Open Sci. 4:170712.

Cornuault J, Delahaie B, Bertrand JAM, Bourgeois YXC, Milaa B, Heeb P, Thalbaud C. 2015. Morphological and plumage colour variation in the union grey white-eye (aves: *Zosterops borbonicus*): assessing the role of selection. Biol J Linn Soc. 114:459–473.

Cronin TW, Marshall NJ. 1989. A retina with at least ten spectral types of photoreceptors in a mantis shrimp. Nature. 339:137–140.

Cuthill IC, Allen WL, Arbuckle K, Caspers B, Chaplin G, Hauber ME, Hill GE, Jablonski NG, Jiggins CD, Kelber A, et al. 2017. The biology of color. Science. 357:eaan0221.

Dalrymple RL, Kemp DJ, Flores-Moreno H, Laffan SW, White TE, Hemmings FA, Tindall ML, Moles AT. 2015. Birds, butterflies and flowers in the tropics are not more colourful than those at higher latitudes. Global Ecol Biogeo. 24:1424–1432.

Dearborn DC, Hanley D, Ballantine K, Cullum J, Reeder DM. 2012. Eggshell colour is more strongly affected by maternal identity than by dietary antioxidants in a captive poultry system. Fun Ecol. 26:912–920.

Delhey K, Delhey V, Kempenaers B, Peters A. 2015. A practical framework to analyze variation in animal colors using visual models. Behav Ecol. 26:367–375.

Delhey K, Peters A. 2008. Quantifying variability of avian colours: are signalling traits more variable? PLoS One. 3:e1689.

Delhey K, Szecsenyi B, Nakagawa S, Peters A. 2017. Conspicuous plumage colours are highly variable. Proc R Soc B. 284:20162593.

Dyer AG, Neumeyer C. 2005. Simultaneous and successive colour discrimination in the honeybee (*Apis mellifera*). J Comp Physiol A Neuroethol Sens Neural Behav Physiol. 191:547–557.

Eaton MD. 2005. Human vision fails to distinguish widespread sexual dichromatism among sexually "monochromatic" birds. Proc Natl Acad Sci USA. 102:10942–10946.

Endler JA, Mielke PW. 2005. Comparing entire colour patterns as birds see them. Biol J Linn Soc. 86:405–431.

Fleishman LJ, Loew ER, Whiting MJ. 2011. High sensitivity to short wavelengths in a lizard and implications for understanding the evolution of visual systems in lizards. Proc Biol Sci. 278:2891–2899.

Garcia JE, Greentree AD, Shrestha M, Dorin A, Dyer AG. 2014. Flower colours through the lens: quantitative measurement with visible and ultraviolet digital photography. PLoS One. 9:e96646.

Garcia JE, Spaethe J, Dyer AG. 2017. The path to colour discrimination is S-shaped: behaviour determines the interpretation of colour models. J Comp Physiol A Neuroethol Sens Neural Behav Physiol. 203:983–997.

Govardovskii VI, Fyhrquist N, Reuter T, Kuzmin DG, Donner K. 2000. In search of the visual pigment template. Vis Neurosci. 17:509–528.

Grether GF, Kolluru GR, Nersissian K. 2004. Individual colour patches as multicomponent signals. Biol Rev Camb Philos Soc. 79:583–610.

Hadfield J. 2010. MCMC methods for multi-response generalized linear mixed models: the mcmcglmm r package. J Stat Softw. 33:1–22.

Hadfield JD, Nakagawa S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol. 23:494–508.

Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. 1998. Multivariate data analysis. Vol. 5. Upper Saddle River (NJ): Pearson Prentice Hall.

Hart NS. 2001. The visual ecology of avian photoreceptors. Prog Retin Eye Res. 20:675–703.

Hempel de Ibarra N, Giurfa M, Vorobyev M. 2001. Detection of coloured patterns by honeybees through chromatic and achromatic cues. J Comp Physiol A. 187:215–224.

Johnsen S. 2016. How to measure color using spectrometers and calibrated photographs. J Exp Biol. 219:772–778.

Johnston G, Lee M, Surasinghe T. 2013. Morphology and allometry suggest multiple origins of rostral appendages in Sri Lankan agamid lizards. J Zool. 289:1–9.

Johnstone RA. 1995. Honest advertisement of multiple qualities using multiple signals. J Theor Biol. 177:87–94.

Kelber A, Vorobyev M, Osorio D. 2003. Animal colour vision — behavioural tests and physiological concepts. Biol Rev. 78:81–118.

Kelber A, Yovanovich C, Olsson P. 2017. Thresholds and noise limitations of colour vision in dim light. Phil Trans R Soc B. 372:20160065.

Kemp DJ, Herberstein ME, Fleishman LJ, Endler JA, Bennett AT, Dyer AG, Hart NS, Marshall J, Whiting MJ. 2015. An integrative framework for the appraisal of coloration in nature. Am Nat. 185:705–724.

Kemp DJ, Rutowski RL. 2011. The role of coloration in mate choice and sexual interactions in butterflies. Adv Study Behav. 43:55–92.

Loew ER, Fleishman LJ, Foster RG, Provencio I. 2002. Visual pigments and oil droplets in diurnal lizards: a comparative study of Caribbean anoles. J Exp Biol. 205:927–938.

Maia R, Eliason CM, Bitton PP, Doucet SM, Shawkey MD. 2013. pavo: an r package for the analysis, visualization and organization of spectral data. Method Ecol Evol. 4:906–913.

Maia R, Rubenstein DR, Shawkey MD. 2016. Selection, constraint, and the evolution of coloration in African starlings. Evolution. 70: 1064–1079.

Maier E. 1992. Spectral sensitivities including the ultraviolet of the passeriform bird *Leiothrix lutea*. J Comp Phys A. 170:709–714.

Marshall J. 2018. Do not be distracted by pretty colors: a comment on Olsson *et al*. Behav Ecol. 29:286–287.

Martin Schaefer H, Schaefer V, Vorobyev M. 2007. Are fruit colors adapted to consumer vision and birds equally efficient in detecting colorful signals? Am Nat. 169(Suppl 1):S159–S169.

Nathans J. 1999. The evolution and physiology of human color vision: insights from molecular genetic studies of visual pigments. Neuron. 24:299–312.

O'Hanlon JC, Holwell GI, Herberstein ME. 2014. Predatory pollinator deception: does the orchid mantis resemble a model species? Curr Zool. 60:90–103.

Oksanen J, Kindt R, Legendre P, Oazhara B, Stevens MHH, Oksanen MJ, Suggests M. 2007. The vegan package. Community Ecology Package. 10:631–637.

Olsson P, Lind O, Kelber A. 2015. Bird colour vision: behavioural thresholds reveal receptor noise. J Exp Biol. 218:184–193.

Olsson P, Lind O, Kelber A. 2018. Chromatic and achromatic vision: parameter choice and limitations for reliable model predictions. Behav Ecol. 29:273–282.

Osorio D, Smith AC, Vorobyev M, Buchanan-Smith HM. 2004. Detection of fruit and the selection of primate visual pigments for color vision. Am Nat. 164:696–708.

Pessoa DM, Maia R, de Albuquerque Ajuz RC, De Moraes PZ, Spyrides MH, Pessoa VF. 2014. The adaptive value of primate color vision for predator detection. Am J Primatol. 76:721–729.

Pike TW. 2012. Preserving perceptual distances in chromaticity diagrams. Behav Ecol. 23:723–728.

Price T, Fialko K. 2018. Receptor noise models: time to consider alternatives?: a comment on Olsson *et al*. Behav Ecol. 29:284–285.

Renoult JP, Kelber A, Schaefer HM. 2017. Colour spaces in ecology and evolutionary biology. Biol Rev Camb Philos Soc. 92:292–315.

Rheindt FE, Fujita MK, Wilton PR, Edwards SV. 2014. Introgression and phenotypic assimilation in *Zimmerius flycatchers* (Tyrannidae): population genetic and phylogenetic inferences from genome-wide SNPs. Syst Biol. 63:134–152.

Schultz TD, Fincke OM. 2013. Lost in the crowd or hidden in the grass: signal apparency of female polymorphic damselflies in alternative habitats. Anim Behav. 86:923–931.

Somaweera R, Somaweera N. 2009. Lizards of Sri Lanka: a colour guide with field keys. Frankfurt am Main (Germany): Edition Chimaira.

Stoddard MC, Prum RO. 2008. Evolution of avian plumage color in a tetrahedral color space: a phylogenetic analysis of new world buntings. Am Nat. 171:755–776.

Stoddard MC, Stevens M. 2011. Avian vision and the evolution of egg color mimicry in the common cuckoo. Evolution. 65:2004–2013.

Troscianko J, Wilson-Aggarwal J, Stevens M, Spottiswoode CN. 2016. Camouflage predicts survival in ground-nesting birds. Sci Rep. 6:19966.

Vasas V, Brebner JS, Chittka L. 2018. Color discrimination is not just limited by photoreceptor noise: a comment on Olsson *et al.* Behav Ecol. 29:285–286.

Vorobyev M, Brandt R, Peitsch D, Laughlin SB, Menzel R. 2001. Colour thresholds and receptor noise: behaviour and physiology compared. Vision Res. 41:639–653.

Vorobyev M, Osorio D. 1998. Receptor noise as a determinant of colour thresholds. Proc Biol Sci. 265:351–358.

Vorobyev M, Osorio D, Bennett AT, Marshall NJ, Cuthill IC. 1998. Tetrachromacy, oil droplets and bird plumage colours. J Comp Physiol A. 183:621–633.

Warton DI, Wright ST, Wang Y. 2012. Distance-based multivariate analyses confound location and dispersion effects. Method Ecol Evol. 3:89–101.

White TE, Dalrymple RL, Herberstein ME, Kemp DJ. 2017. The perceptual similarity of orb-spider prey lures and flower colours. Evol Ecol. 31:1–20.

White TE, Kemp DJ. 2017. Colour polymorphic lures exploit innate preferences for spectral versus luminance cues in dipteran prey. BMC Evol Biol. 17:191.

Whiting MJ, Noble DW, Somaweera R. 2015. Sexual dimorphism in conspicuousness and ornamentation in the enigmatic leaf-nosed lizard *Ceratophora tennentii* from Sri Lanka. Biol J Linn Soc. 116:614–625.