RESEARCH ARTICLE

Running head: COMPARING COLOURS

Comparing colours using visual models

Rafael Maia[1,*] & Thomas E. White[2,†]

[1] Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, NY 10027, U.S.A.

[2] Department of Biological Science, Macquarie University, North Ryde, 2109, Australia

[*] rm3368@columbia.edu

[†] thomas.white@mq.edu.au

Keywords: vision, dimorphism, polymorphism, mimicry, crypsis, multivariate statistics

Abstract

1. Colour 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 perceptual colour spaces, which allow for the viewer-subjective estimation of colour appearance. Most studies of colour in camouflage, aposematism, sexual selection, and other signalling contexts draw on these colour spaces, with the shared analytical objective of estimating how similar (or dissimilar) colour samples are to a given viewer.

2. We summarise popular approaches for estimating the separation of samples in colour 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 colour samples by neglecting (i) the statistical distribution and within-group variation of the data, and/or (ii) the perceptual separation of groups relative to the observer's visual capabilities.

3. Instead, we formalize the two questions that must be answered to establish both the statistical presence and perceptual magnitude of colour differences, and propose a two-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 colour-data features such as heterogeneity and outliers.

4. We demonstrate the pitfalls of current methods and the flexibility of our suggested framework using a example from the literature, with recommendations for future inquiry.

Introduction

Humans, as most primates, are an extremely visually-oriented species (Vorobyev, 2004), and the study of colour in nature has driven fundamental advances in ecology and evolutionary biology (Cuthill *et al.*, 2017). Colour is a subjective experience, however, so substantial effort has been dedicated to measuring colours "objectively" (Garcia *et al.*, 2014; Johnsen, 2016) through visual models to consider the perspective of ecologically relevant viewers (Kemp *et al.*, 2015; Renoult *et al.*, 2017). These models have radically advanced the study of colour traits by allowing researchers to account for the factors influencing the generation and perception 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; Endler & Mielke, 2005; Kelber *et al.*, 2003; Vorobyev & Osorio, 1998).

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

This colour space representation is convenient for several reasons. It offers an intuitive way of analysing phenotypes that we cannot measure directly: we can estimate how animals with different visual systems "see" different colours by representing them in a Cartesian coordinate system, producing a receiver-dependent morphospace (Kelber *et al.*, 2003; Renoult *et al.*, 2017). Further, it allows estimating how similar or dissimilar colours are *to a given observer*, by measuring the distance between colour points in its colour space (Endler & Mielke, 2005; Vorobyev *et al.*, 1998; Vorobyev & Osorio, 1998). Finally, we can integrate behavioural and psychophysical data into these models to predict whether an observer could effec-

tively discriminate pairs of colours (Chittka, 1992; Vorobyev *et al.*, 2001; Vorobyev & 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 & Fincke, 2013; Whiting *et al.*, 2015), the precision of mimicry (O'Hanlon *et al.*, 2014; White *et al.*, 2017), the extent of signal variability among populations or species (Delhey & Peters, 2008; Rheindt *et al.*, 2014), or the effect of experimental treatments (Barry *et al.*, 2015). At the heart of these inquiries lies the same question: *how different are these colours to the animal viewing them?*. While a further distinction is often drawn between questions dealing with colours that are very similar ('discriminability' near-threshold) and very different ('perceptual distance' supra-threshold; Kemp *et al.*, 2015), accruing empirical evidence suggests this is largely artificial (Fleishman *et al.*, 2016; Kemp *et al.*, 2015; van der Kooi *et al.*, 2016). We thus refer to both as "questions of discriminability", and while the discussion below largely centres on near-threshold scenarios, the presented methods are broadly applicable.

Challenges in estimating the discriminability of colour samples

The receptor noise-limited model of Vorobyev & Osorio (1998) has proven particularly useful for addressing questions of discriminability. The model assumes that that chromatic and achromatic channels operate independently and that the limits to colour discrimination are set by noise arising in receptors and subsequent neural processing (Vorobyev *et al.*, 1998; Vorobyev & Osorio, 1998). This noise is dependent on the receptor type and abundance on the retina, as well Weber's law ($k = \Delta I / I$) more generally, which ultimately establishes the unit of Just Noticeable Differences (JND Vorobyev *et al.*, 2001). It follows that values lower than 1 JND ($\frac{signal}{noise} < 1$) are predicted to be indistinguishable, while values greatly above this threshold are ubiquitously distinct. This provides a useful standard for estimating the similarity of groups of points in colour space: the greater the distance between

colours, 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, crypsis ineffective. This offers a clear link between variation and classification within a perceptual framework, and has been widely used for this purpose (Barry *et al.*, 2015; Delhey & Peters, 2008; O'Hanlon *et al.*, 2014; Schultz & Fincke, 2013; White *et al.*, 2017).

To adequately compare samples of colours, however, it is necessary to determine if the average distance between them is both statistically and perceptually meaningful (i.e. above-threshold; Endler & Mielke, 2005). Commonly, an "average colour" for each group is derived by taking a mean reflectance spectrum or by averaging their position in colour space. In either case, the colour distance between groups is then calculated from these mean quantum catches per-receptor per-group — their centroids in multivariate space (Fig. 1, bold arrow). However, the centroid obtained from arithmetic means of receptor coordinates is not an appropriate measure of location for this goal, since colour distances are perceived in a ratio scale (Cardoso & Gomes, 2015). Instead, the geometric mean must be used. Further, 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 & Shultz, 2012; Maia *et al.*, 2016).

An alternative approach calculates the pairwise distances between all points in group A and group B, then averaging these distances to obtain a mean distance between groups (Fig. 1, thin arrows; e.g. Barry *et al.*, 2015; Dearborn *et al.*, 2012). 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 the threshold. The average linkage, however, is also inadequate because it conflates within- and among-group variation. This is because Euclidean distances (and by extension JNDs) are *translation-invariant*: they ignore the position of points

in colour space and the direction of the distance vectors, reflecting only the magnitude of differences between two points. Therefore, the average linkage reduces to a measure of spread, and will scale with both within- and between-group distances 1.

As these issues show, hypotheses of discriminability have primarily focused on testing whether the difference between samples is above a perceptual threshold. However, the convenience of such thresholds belies fact that simply comparing means between groups is not sufficient to infer, statistically, whether samples are different. To answer if two groups are different, one must compare the variation between- and within-groups. This is particularly problematic in the case of colours that function as signals in social interactions (e.g. Kemp & 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 colour 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 colour in mate choice, but is not sexually dichromatic (Fig. 1). In this species colour 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 1JND, despite them being sampled from the same (statistical) population.

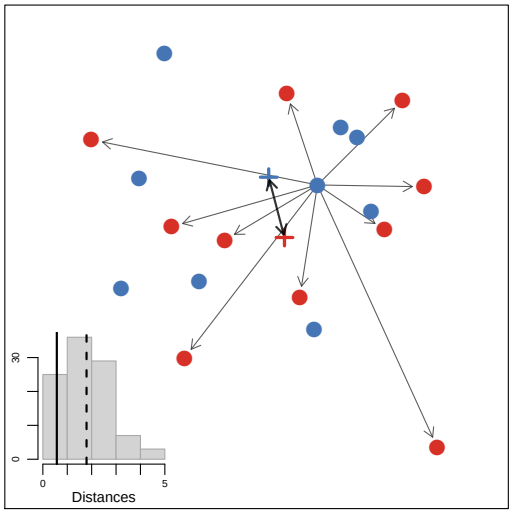

Figure 1: The link distance (i.e. average pairwise distance between groups) conflates within- and among-group variation. Here, two samples were drawn from the same distribution. Thin arrows represent distances between a random point in the first sample (blue) and all points from the second sample (red), all of which are greater than the distance between the geometric means of the two 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).

The limitations of current methods for comparing colour space distributions

Several methods have been proposed to subvert the aforementioned issues by accounting for the relative distributions of samples in colour 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, 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 & Prum (2008) is the volume overlap. In this approach, the volume occupied by a sample of colours is estimated from its enveloping convex hull, and separation between samples is inferred from their overlap.Stoddard & Stevens (2011) used this metric to show that a

greater overlap in colour volume between cuckoo and host eggs is associated with lower rejection of parasitic eggs. This approach is appealing because it considers the distribution of colour points in multivariate space, though there are limits to its interpretation: (i) 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) and (ii) 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 colour 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 colours was proposed by Endler & Mielke (2005), who devised a non-parametric 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 & Mielke, 2005). Like other distance-based methods, it is sensitive to confounding heterogeneity among samples when testing for differences in *location*(Anderson & Walsh, 2013; Warton *et al.*, 2012). Despite its 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 two hypotheses that are sel-

dom formally distinguished: (i) that the focal samples are statistically distinct, and (ii) 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 variation or uncertainty in estimates. We demonstrate these issues using a simulation-based approach by testing the efficacy of popular methods in detecting the separation of groups in colour 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 colour space, we simulated data analogous to that obtained from applying an avian visual model to spectral reflectance data. Birds are tetrachromatic (Hart, 2001) and colours will thus be represented by the quantum catches of its four photoreceptors (though the procedure followed here can be applied to visual systems with any number of receptors). For each replicate, we simulated two samples defined by four variables (*usml* photoreceptors) taken from log-normal distributions (since quantum catches are non-negative and perceptual distances follow a ratio scale, as defined by the Weber fraction, described above). We generated samples following two 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 perceptually distinct.

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 $\mathcal{U}(0, 10)$, and standard deviation proportional to the mean: $\sigma_i = a_i \mu_{iA}$, with $a_i \sim \mathcal{U}(0, 0.5)$

(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 two samples with varying degrees of separation proportional to the within-group variance, we used a multivariate effect size S obtained by calculating a constant $k_i = \frac{S}{\sqrt{n}}\bar{\sigma}_i$, 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 $\mu_{iB} = \mu_{iA} + k_i$ and σ_i. Thus, our simulations effectively produced two samples with Mahalanobis Distance $D_M \sim S$. We simulated data for $S = \{0, 0.1, 0.25, 0.5, 0.75, 1.0, 1.5, 2, 2.5, 3.0\}$ (Fig. 2), replicated 200 times for sample sizes $N = \{10, 20, 50, 100\}$ each.

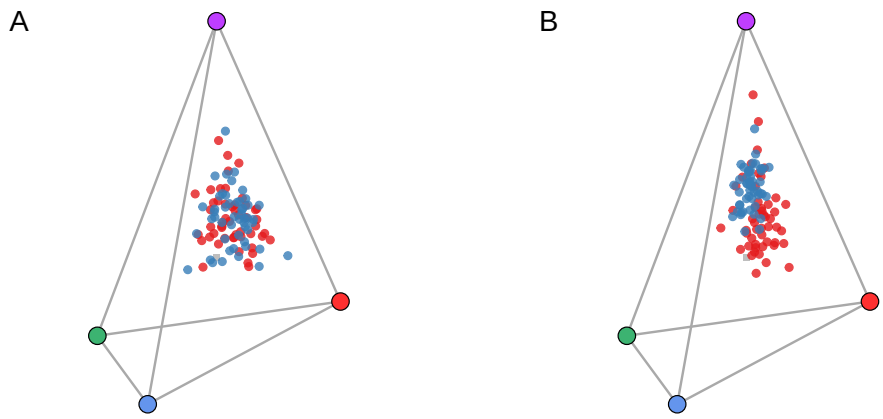

Figure 2: Example simulated data for the two groups (red, blue) in a tetrahedral colourspace. Shown here are data with sample size $N = 50$ and effect size (A) $S = 0$ and (B) $S = 3$.

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 $\mu_{iA} \sim \mathcal{U}(0, 10)$, 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 $\mu_{iB} = k_i \mu_{iA}$, where $k_i \sim \mathcal{U}(0.88, 1.12)$, resulting in an average distance between geometric means (hereafter, "mean distance") of 1.11 (95% quantiles: $0.35 - 2.77$) and within-group average pairwise distance of 4.46 (95% quantiles: $1.03 - 11.10$) after 1000 replicates.

After the two groups were simulated, we used the `R` package `pavo` (Maia *et al.*, 2013) to calculate colour distances using relative receptor densities values 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 distances Delhey & Peters (2008); Pike (2012) 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 for extensions of this approach). Third, we calculated the volume overlap between the two samples (relative to their combined volumes) in a tetrahedral colour space defined by the receptors' relative quantum catches (thus disregarding receptor noise; Stoddard & Prum, 2008). Finally, we calculated the volume overlap for the XYZ Cartesian coordinates based on perceptual distances, generating a colour 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 (Fig. 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 (Fig. 3,4).

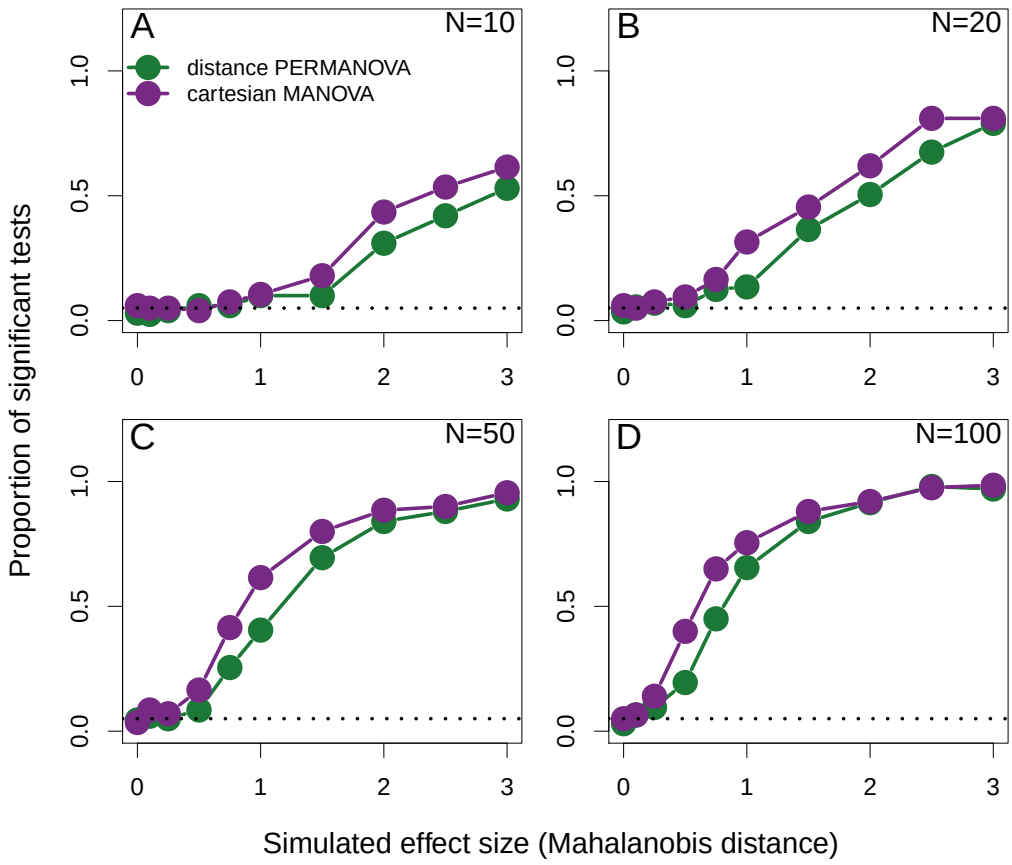

Figure 3: Power and Type I error rate of the distance PERMANOVA (green) and Cartesian MANOVA (purple). Panels show the proportion of simulations yielding significant results for each approach under different sample and effect sizes.

The two approaches showed some disagreement, with between $10 - 15\%$ of

the simulations significant only in one of the two tests (Fig.4). This disagreement was not random, being concentrated at smaller effect sizes with increasing sample sizes, and also with the Cartesian MANOVA being more likely to be significant when the distance PERMANOVA was not than vice-versa, at an approximately constant rate across sample sizes (Fig.4).

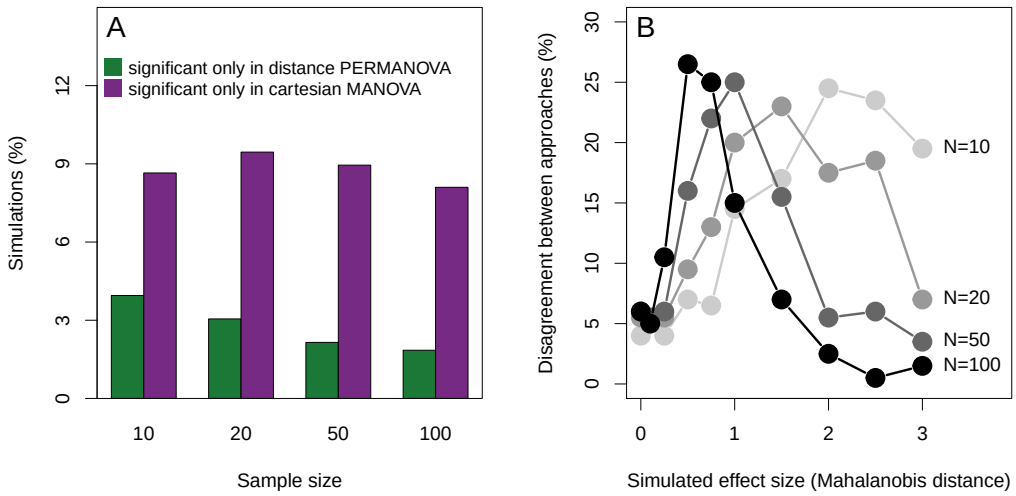

Figure 4: The disagreement between multivariate statistical approaches when testing for separation between samples in colour space.

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 $1JND$ mark (Fig. 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 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).

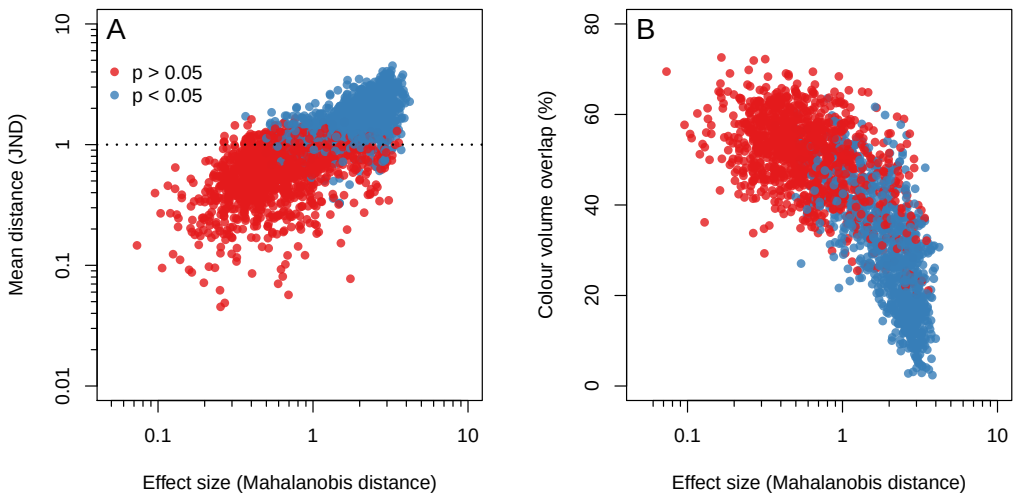

Figure 5: The association between effect size and (A) mean distance and (B) colour volume overlap. Significant distance PERMANOVA results are in blue, whereas non-significant results are in red. Dotted line indicates the threshold of 1 JND.

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. Simulations produced a wide range of outcomes, with non-significant and significant tests both above and below the perceptual threshold of 1JND (Fig. 6). In contrast with the power simulations above (Fig 5), the significance threshold did not match the perceptual threshold. As in the hypothetical example from the introduction, some simulated data were statistically indistinguishable despite having mean distances above the perceptual threshold (Fig. 6, dark red). Likewise, other simulations produced samples that were statistically different, but that difference was below the perceptual threshold and therefore undetectable to its observer (Fig. 6, dark blue points). These results highlight the importance of considering both statistical separation and perceptual thresholds when testing the hypothesis that samples are perceptually discriminable.

Figure 6A shows that, intuitively, tests were significant when within-group differences were small relative to among-group 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 0.5JND smaller than the within-group average link distance (Fig. 6A, grey 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% (Fig. 6B, horizontal line at 3%).

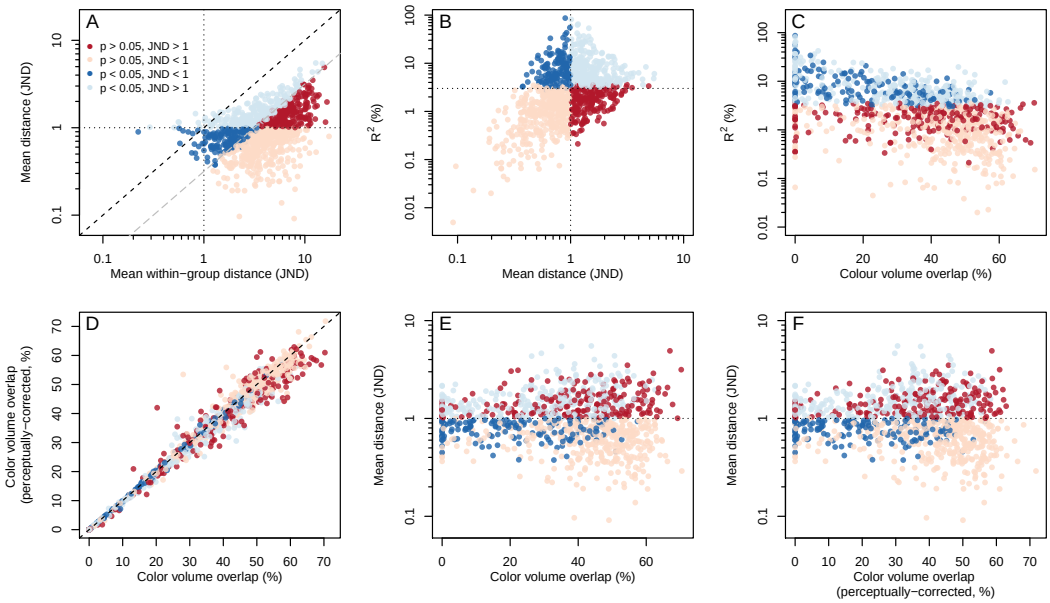

Figure 6: Results from threshold simulation. Red and blue denote non-significant and significant PERMANOVA tests, respectively, and light colours denote when that approach would yield the same inference as comparing mean distances to a threshold of 1JND. Thus, dark blue points indicate a significant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

Though there is a negative association between R^2 and volume overlap, 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 (Fig. 6C, E-F). In other words, even complete

separation in colour volumes can result in non-significant, below-threshold cases, since samples can be contiguous in perceptual space (Fig. 6E-F). Likewise, samples can have high overlap but be entirely distinguishable statistically and perceptually. Further, there is no association between volume overlap and mean distance between groups (Fig. 6E,F). Importantly, these results were unaltered by considering receptor noise in the volume overlap calculation, since these are still strongly and positively correlated with their non-perceptual counterparts (Fig. 6D,F).

A two-step approach to estimate statistical and perceptual separation

As described above, questions of discriminability require testing two distinct hypotheses: if samples are (i) statistically and (ii) perceptually distinct. We therefore propose a two-step answer to such questions, which explicitly formalizes these hypotheses. For the first question — are the samples *statistically separate* in colour space? — we show that both a PERMANOVA using perceptual colour distances (Anderson, 2005; Cornuault *et al.*, 2015), and a MANOVA using perceptually-calibrated Cartesian coordinates (Delhey & Peters, 2008; Pike, 2012) are well suited. Both exclude achromatic variation and properly account for the multivariate nature of the data. There are also minimal discrepancy between them (Fig. 3,4), so the decision between them may be informed by convenience and the structure of the data at hand. The conceptual simplicity and non-parametric robustness of the PERMANOVA has seen it widely adopted in community ecology and genomics, and it is the least sensitive distance-based non-parametric approach to within-group dispersion and correlation structure heterogeneity (Anderson & Walsh 2013; though see Warton *et al.* 2012 for broader limitations of distance-based methods, which are applicable to colour data; Endler & Mielke 2005).

Once the separation of samples is established statistically, a second question must be answered: is this separation *perceptually discriminable*? 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 es-

timating the distance in colour space between group geometric means rather than through the average pairwise distance or volume-overlap based metrics, which fail to accurately estimate group separation (Figs. 1,6). 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 re-sampling (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 perceptual threshold of adequate biological significance, one can conclude that the samples being compared are distinct and discriminable.

Empirical example: Sexual dichromatism in the leaf-nosed lizard *Ceratophora tennentii*

Visually signalling animals often use distinct body parts for different purposes, such as social signalling to mates or warning predators (Barry *et al.*, 2015; Grether *et al.*, 2004; Johnstone, 1995). The nature of intraspecific variation in colour can thus inform on 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 colour pattern that is reduced or absent in the other (Bell & Zamudio, 2012; Kemp & Rutowski, 2011).

Dragon lizards (Agamidae) are known for variable colouration used in both social and anti-predator contexts (Johnston *et al.*, 2013; Somaweera & Somaweera, 2009). The leaf-nosed lizard *Ceratophora tennentii* has multiple discrete colour patches, with apparent sex differences between body parts (Fig. 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.

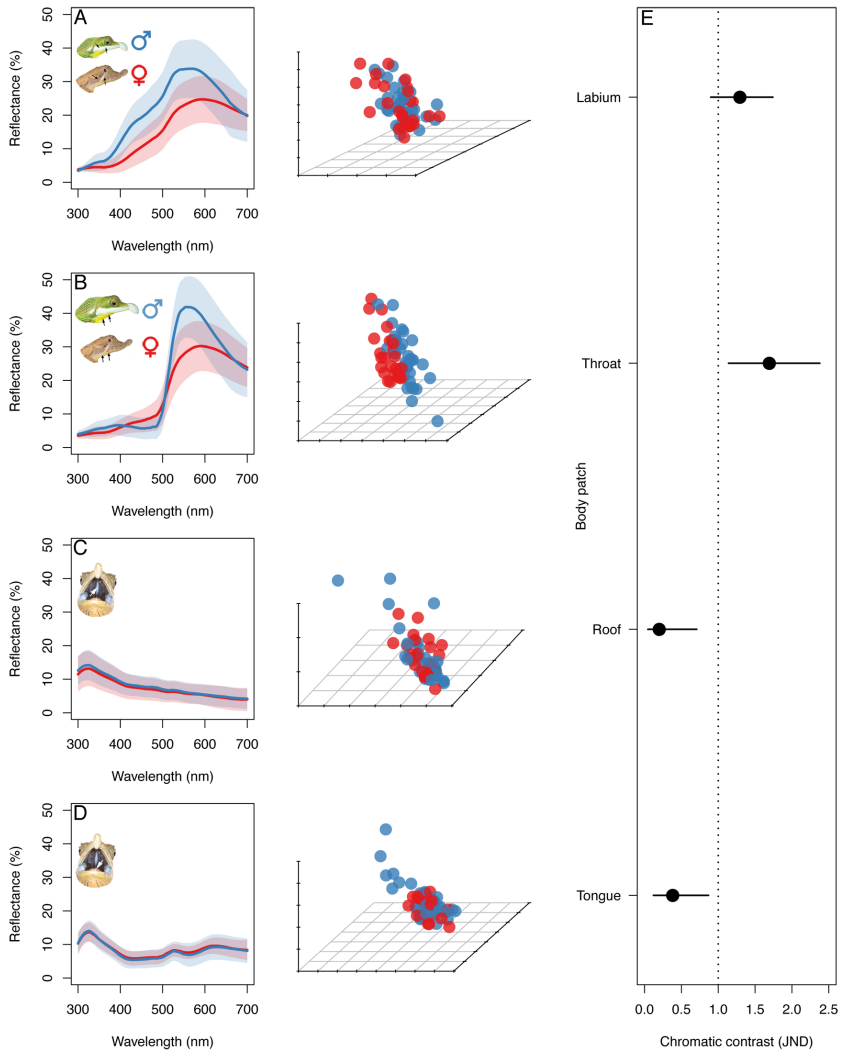

Figure 7: The mean (± SD) spectral reflectance of female (red) and male (black) (A) labial, (B) throat, (C) mouth-roof, and (D) tongue (left panels), and their colourspace distribution according in a tetrachromatic model of agamid vision (middle panels). Inset images indicate approximate sampling regions.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

Following standard calculations for the log-linear receptor-noise model, we used the spectral sensitivity of *Ctenophorus ornatus* ($\lambda_{max} = 360, 440, 493, 571$ nm) as modelled according to a vitamin A1 template (Barbour *et al.*, 2002; Govardovskii *et al.*, 2000). We assumed a relative photoreceptor density of 1 : 1 : 3.5 : 6, and a

signal-to-noise ratio yielding a Weber fraction of 0.1 for the long-wavelength cone (Fleishman *et al.*, 2011; Vorobyev & Osorio, 1998). We tested each body region separately using PERMANOVA. As above, we used the `R` package `pavo` for visual modelling, 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 ($F_{1,57} = 13.96$, $P < 0.01$; Fig. 7A,B), but not for tongues ($F_{1,58} = 1.63$, $P = 0.22$) or mouth-roofs ($F_{1,55} = 0.52$, $P = 0.50$; Fig. 7C,D). However, bootstraps of group separation suggest that intersexual differences in labial colour are likely imperceptible to conspecifics (Fig. 7E). Our results therefore suggest the absence of dichromatism in all but throat colour from the lizard perspective, despite statistical significance for the labial region. These results do not thus implicate sexual selection as a strong driver of intersexual colour differences in these body regions of *C. ornatus*.

Discussion

Visual models offer a powerful tool for quantifying the subjective perception of colour, which—as the ultimate canvas for colour-signal evolution—offers direct 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. Here, we show that most methods typically fail to consider these aspects, proposing a flexible alternative that explicitly addresses both.

The use of visual models that do not consider the role of receptor noise in colour discrimination is often justified on the basis of simplifying and relaxing assumptions about colour perception, since intricate empirical work is necessary to estimate parameters necessary for these models with precision(Kelber *et al.*, 2017; Olsson *et al.*, 2015; Vorobyev & Osorio, 1998). However, we contend that, on the contrary, simplified models may actually make very strong latent assumptions,

which are not supported by the empirical evidence. Namely, "simpler" models generally implicitly assume that all cones contribute equally to colour perception, that colour discrimination is unequivocal (i.e. there is no threshold of detectability) and that colour differences follow an interval scale (as opposed to a ratio scale). Thus, we argue that detectability relative to a threshold is essential for tests of discriminability.

Our simulations show that both the distance PERMANOVA and Cartesian MANOVA perform similarly well in statistically differentiating colours in perceptual space (Fig. 3). Studies have pointed out that distance-based methods perform poorly when the experimental design is unbalanced or when there is heteroscedasticity (Anderson & Walsh, 2013; Warton *et al.*, 2012). However, this might still be the most robust option for high-dimensional visual systems (e.g. Arikawa *et al.*, 1987; Cronin & Marshall, 1989), reducing data to a single metric of distance. Recently, Delhey & Peters (2008) advocated a similar MANOVA approach, by applying a Principal Component Analysis (PCA) to the perceptually-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). While we have focused on tests of differences in the location of colours in colour space, we recognise that other characteristics — such as differences in dispersion and correlation structure — might themselves be of biological interest, for which a PCA approach may be particularly useful.

The MANOVA approach can be extended to multivariate generalizations of generalized linear models using the perceptually-corrected Cartesian coordinates as response variables (Hadfield, 2010). These models can also relax the assumptions of heteroscedasticity by estimating the variance-covariance of the responses (Hadfield, 2010), and can be extended to include various error and model structures, such as hierarchical and phylogenetic models (Hadfield & Nakagawa, 2010).

Still, these approaches will only test for the statistical separation in colour space, so the magnitude of that separation compared to the threshold 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). In either case, it is essential to parametrize the underlying visual model appropriately. The Weber fraction and receptor densities chosen will strongly affect perceptual distances(Bitton *et al.*, 2017) since they directly scale with the JND unit. 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; Schaefer *et al.*, 2007).

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

Implementation and data accessibility

Analyses and simulations can be found in `https://github.com/rmaia/msdichromatism/`, and will be implemented in the `R` package `pavo` release accompanying this manuscript. They are temporarily available from the repository (`/R/` folder). The function `bootcentroiddS` calculates the bootstrapped confidence intervals for mean dis-

tances, and `jnd2xyz` converts chromatic distances in JNDs to Cartesian coordinates. These functions require the bleeding edge version of `pavo`, which can be found at https://github.com/rmaia/pavo/. Lizard colour data (Whiting *et al.*, 2015) is available at http://dx.doi.org/10.6084/m9.figshare.1452908.

Acknowledgments

We would like to thank Ruchira Somaweera for leaf-nosed lizard photographs, and Dan Noble 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

RM and TEW conceived the ideas, designed methodology, analysed the data, and wrote the manuscript.

References

Anderson, M.J. (2005) Permutational multivariate analysis of variance. *Department of Statistics, University of Auckland, Auckland*, **26**, 32–46.

Anderson, M.J. & Walsh, D.C. (2013) Permanova, anosim, and the mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? *Ecological Monographs*, **83**(4), 557–574.

Arikawa, K., Inokuma, K. & Eguchi, E. (1987) Pentachromatic visual system in a butterfly. *Naturwissenschaften*, **74**(6), 297–298.

Avilés, J.M., Soler, J.J. & Hart, N.S. (2011) Sexual selection based on egg colour: physiological models and egg discrimination experiments in a cavity-nesting bird. *Behavioral Ecology and Sociobiology*, **65**(9), 1721–1730.

Barbour, H.R., Archer, M.A., Hart, N.S., Thomas, N., Dunlop, S.A., Beazley, L.D. & Shand, J. (2002) Retinal characteristics of the ornate dragon lizard, ctenophorus ornatus. *Journal of Comparative Neurology*, **450**(4), 334–344.

Barry, K.L., White, T.E., Rathnayake, D.N., Fabricant, S.A. & Herberstein, M.E. (2015) Sexual signals for the colour-blind: cryptic female mantids signal quality through brightness. *Functional Ecology*, **29**(4), 531–539.

Bell, R.C. & Zamudio, K.R. (2012) Sexual dichromatism in frogs: natural selection, sexual selection and unexpected diversity. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20121609.

Bitton, P.P., Janisse, K. & Doucet, S.M. (2017) Assessing sexual dicromatism: The importance of proper parameterization in tetrachromatic visual models. *PLOS ONE*, **12**(1), 1–22.

Blonder, B., Babich Morrow, C., Maitner, B., Harris, D.J., Lamanna, C., Violle, C., Enquist, B.J. & Kerkhoff, A.J. (2017) New approaches for delineating n-dimensional hypervolumes. *Methods in Ecology and Evolution*, n/a–n/a.

Burns, K.J. & Shultz, A.J. (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**(2), 211–221.

Cardoso, G.C. & Gomes, A.C.R. (2015) Using reflectance ratios to study animal coloration. *Evolutionary Biology*, **42**(3), 387–394.

Chittka, L. (1992) The colour hexagon: a chromaticity diagram based on photoreceptor excitations as a generalized representation of colour opponency. *Journal of Comparative Physiology A*, **170**(5), 533–543.

Cornuault, J., Delahaie, B., Bertrand, J.A.M., Bourgeois, Y.X.C., MilÃą, B., Heeb, P. & ThÃľbaud, C. (2015) Morphological and plumage colour variation in the rÃľunion grey white-eye (aves: Zosterops borbonicus): assessing the role of selection. *Biological Journal of the Linnean Society*, **114**(2), 459–473.

Cronin, T.W. & Marshall, N.J. (1989) A retina with at least ten spectral types of photoreceptors in a mantis shrimp. *Nature*, **339**(6220), 137–140.

Cuthill, I.C., Allen, W.L., Arbuckle, K., Caspers, B., Chaplin, G., Hauber, M.E., Hill, G.E., Jablonski, N.G., Jiggins, C.D., Kelber, A., Mappes, J., Marshall, J., Merrill, R., Osorio, D., Prum, R., Roberts, N.W., Roulin, A., Rowland, H.M., Sherratt, T.N., Skelhorn, J., Speed, M.P., Stevens, M., Stoddard, M.C., Stuart-Fox, D., Talas, L., Tibbetts, E. & Caro, T. (2017) The biology of color. *Science*, **357**(6350).

Dearborn, D.C., Hanley, D., Ballantine, K., Cullum, J. & Reeder, D.M. (2012) Eggshell colour is more strongly affected by maternal identity than by dietary antioxidants in a captive poultry system. *Functional Ecology*, **26**(4), 912–920.

Delhey, K. & Peters, A. (2008) Quantifying variability of avian colours: are signalling traits more variable? *PLoS One*, **3**(2), e1689.

Delhey, K., Szecsenyi, B., Nakagawa, S. & Peters, A. (2017) Conspicuous plumage

colours are highly variable. *Proceedings of the Royal Society of London B: Biological Sciences*, **284**(1847).

Eaton, M.D. (2005) Human vision fails to distinguish widespread sexual dichromatism among sexually "monochromatic" birds. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(31), 10942–10946.

Endler, J.A. & Mielke, P.W. (2005) Comparing entire colour patterns as birds see them. *Biological Journal of the Linnean Society*, **86**(4), 405–431.

Fleishman, L.J., Loew, E.R. & Whiting, M.J. (2011) High sensitivity to short wavelengths in a lizard and implications for understanding the evolution of visual systems in lizards. *Proceedings of the Royal Society of London B: Biological Sciences*, **278**(1720), 2891–2899.

Fleishman, L.J., Perez, C.W., Yeo, A.I., Cummings, K.J., Dick, S. & Almonte, E. (2016) Perceptual distance between colored stimuli in the lizard anolis sagrei: comparing visual system models to empirical results. *Behavioral ecology and sociobiology*, **70**(4), 541–555.

Garcia, J.E., Greentree, A.D., Shrestha, M., Dorin, A. & Dyer, A.G. (2014) Flower colours through the lens: Quantitative measurement with visible and ultraviolet digital photography. *PloS one*, **9**(5), e96646.

Govardovskii, V.I., Fyhrquist, N., Reuter, T., Kuzmin, D.G. & Donner, K. (2000) In search of the visual pigment template. *Visual neuroscience*, **17**(04), 509–528.

Grether, G.F., Kolluru, G.R. & Nersissian, K. (2004) Individual colour patches as multicomponent signals. *Biological Reviews*, **79**(03), 583–610.

Hadfield, J.D. & Nakagawa, S. (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, **23**(3), 494–508.

Hadfield, J. (2010) Mcmc methods for multi-response generalized linear mixed models: The mcmcglmm r package. *Journal of Statistical Software, Articles*, **33**(2), 1–22, doi:10.18637/jss.v033.i02, URL https://www.jstatsoft.org/v033/i02.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L. *et al.* (1998) *Multivariate data analysis*, vol. 5. Prentice hall Upper Saddle River, NJ.

Hart, N.S. (2001) The visual ecology of avian photoreceptors. *Progress in retinal and eye research*, **20**(5), 675–703.

Johnsen, S. (2016) How to measure color using spectrometers and calibrated photographs. *Journal of Experimental Biology*, **219**(6), 772–778.

Johnston, G., Lee, M. & Surasinghe, T. (2013) Morphology and allometry suggest multiple origins of rostral appendages in sri lankan agamid lizards. *Journal of Zoology*, **289**(1), 1–9.

Johnstone, R.A. (1995) Honest advertisement of multiple qualities using multiple signals. *Journal of theoretical Biology*, **177**(1), 87–94.

Kelber, A., Vorobyev, M. & Osorio, D. (2003) Animal colour visionâĂŞbehavioural tests and physiological concepts. *Biological Reviews*, **78**(01), 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**(1717), 20160065.

Kemp, D.J., Herberstein, M.E., Fleishman, L.J., Endler, J.A., Bennett, A.T.D., Dyer, A.G., Hart, N.S., Marshall, J. & Whiting, M.J. (2015) An integrative framework for the appraisal of coloration in nature. *The American Naturalist*, **185**, 705–724.

Kemp, D.J. & Rutowski, R.L. (2011) The role of coloration in mate choice and sexual interactions in butterflies. *Advances in the Study of Behavior*, **43**(5).

Maia, R., Eliason, C.M., Bitton, P.P., Doucet, S.M. & Shawkey, M.D. (2013) pavo: an r package for the analysis, visualization and organization of spectral data. *Methods in Ecology and Evolution*, 906–913.

Maia, R., Rubenstein, D.R. & Shawkey, M.D. (2016) Selection, constraint, and the evolution of coloration in african starlings. *Evolution*, **70**(5), 1064–1079, doi:10.1111/evo.12912, URL http://dx.doi.org/10.1111/evo.12912.

O'Hanlon, J.C., Holwell, G.I. & Herberstein, M.E. (2014) Predatory pollinator deception: Does the orchid mantis resemble a model species? *Current Zoology*, **60**(1).

Oksanen, J., Kindt, R., Legendre, P., OâĂŹHara, B., Stevens, M.H.H., Oksanen, M.J. & 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. *Journal of Experimental Biology*, **218**(2), 184–193.

Osorio, D., Smith, A.C., Vorobyev, M. & Buchanan-Smith, H.M. (2004) Detection of fruit and the selection of primate visual pigments for color vision. *The American Naturalist*, **164**(6), 696–708.

Pessoa, D.M.A., Maia, R., de Albuquerque Ajuz, R.C., De Moraes, P.Z.P.M., Spyrides, M.H.C., Pessoa, V.F. *et al.* (2014) The adaptive value of primate color vision for predator detection. *American journal of primatology*, **76**(8), 721–729.

Pike, T.W. (2012) Preserving perceptual distances in chromaticity diagrams. *Behavioral Ecology*, **23**(4), 723–728.

Renoult, J.P., Kelber, A. & Schaefer, H.M. (2017) Colour spaces in ecology and evolutionary biology. *Biological Reviews*, **92**(1), 292–315.

Rheindt, F.E., Fujita, M.K., Wilton, P.R. & Edwards, S.V. (2014) Introgression and phenotypic assimilation in zimmerius flycatchers (tyrannidae): Population genetic and phylogenetic inferences from genome-wide snps. *Systematic Biology*, **63**(2), 134–152.

Schaefer, H., Schaefer, V. & Vorobyev, M. (2007) Are fruit colors adapted to consumer vision and birds equally efficient in detecting colorful signals? *The American Naturalist*, **169**(S1), S159–S169. PMID: 19426090.

Schultz, T.D. & Fincke, O.M. (2013) Lost in the crowd or hidden in the grass: signal apparency of female polymorphic damselflies in alternative habitats. *Animal behaviour*, **86**(5), 923–931.

Somaweera, R. & Somaweera, N. (2009) *Lizards of Sri Lanka: a colour guide with field keys*. Andreas S. Brahm.

Stoddard, M.C. & Prum, R.O. (2008) Evolution of avian plumage color in a tetrahedral color space: a phylogenetic analysis of new world buntings. *The American Naturalist*, **171**(6), 755–776.

Stoddard, M.C. & Stevens, M. (2011) Avian vision and the evolution of egg color mimicry in the common cuckoo. *Evolution*, **65**(7), 2004–2013.

Troscianko, J., Wilson-Aggarwal, J., Stevens, M. & Spottiswoode, C.N. (2016) Camouflage predicts survival in ground-nesting birds. *Scientific reports*, **6**.

van der Kooi, C.J., Elzenga, J.T.M., Staal, M. & Stavenga, D.G. (2016) How to colour a flower: on the optical principles of flower coloration. *Proc. R. Soc. B*, vol. 283. The Royal Society, vol. 283, p. 20160429.

Vorobyev, M., Osorio, D., Bennett, A.T., Marshall, N. & Cuthill, I. (1998) Tetrachromacy, oil droplets and bird plumage colours. *Journal of Comparative Physiology A*, **183**(5), 621–633.

Vorobyev, M. (2004) Ecology and evolution of primate colour vision. *Clinical and Experimental Optometry*, **87**(4-5), 230–238.

Vorobyev, M., Brandt, R., Peitsch, D., Laughlin, S.B. & Menzel, R. (2001) Colour thresholds and receptor noise: behaviour and physiology compared. *Vision research*, **41**(5), 639–653.

Vorobyev, M. & Osorio, D. (1998) Receptor noise as a determinant of colour thresholds. *Proceedings of the Royal Society of London B: Biological Sciences*, **265**(1394), 351–358.

Warton, D.I., Wright, S.T. & Wang, Y. (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, **3**(1), 89–101.

White, T.E., Dalrymple, R.L., Herberstein, M.E. & Kemp, D.J. (2017) The perceptual similarity of orb-spider prey lures and flower colours. *Evolutionary Ecology*, **31**(1), 1–20.

Whiting, M.J., Noble, D.W. & Somaweera, R. (2015) Sexual dimorphism in conspicuousness and ornamentation in the enigmatic leaf-nosed lizard ceratophora tennentii from sri lanka. *Biological Journal of the Linnean Society*, **116**(3), 614–625.