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 heuristic examples drawn 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 in order to explicitly 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 receiver visual systems (Chittka, 1992; Endler & Mielke, 2005; Kelber *et al.*, 2003; Vorobyev & Osorio, 1998).

Several forms of visual models exist and are currently used, which vary in their assumptions about the psychophysical properties of visual systems and visual processing (Chittka, 1992; Endler & Mielke, 2005; Vorobyev & Osorio, 1998). Despite this variation, all models invariably attempt to delimit a colour space informed by the number and sensitivity of photoreceptors in an animal's retina (Renoult *et al.*, 2017). Individual colours can then be represented in this space, with their location being determined by the degree to which reflected light differentially stimulates the viewers' receptors. Our own trichromatic vision, for example, can be represented by a triangle with the "red", "green", and "blue" cones as its vertices. Any coloured stimulus can then appear as a point in this space, such that a colour that exclusively triggers one of these receptors will fall on that vertex, while a white or black (i.e. achromatic) colour that stimulates all cones equally will lie at the geometric centre.

Representing colours in dedicated spaces is convenient for several reasons. First, it offers an intuitive way of analysing phenotypes that we cannot measure directly: we can instead estimate how animals with different visual systems "see" different colours by representing them in a Cartesian coordinate system, thereby producing a receiver-dependent morphospace (Kelber *et al.*, 2003; Renoult *et al.*, 2017). Second, it allows for the estimation of how similar or dissimilar colours are *to a given observer*, by estimating 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 models in order to predict whether an observer could effectively discriminate pairs of colours, or if they would instead be perceptually indistinguishable (Chittka, 1992; Vorobyev *et al.*, 2001; Vorobyev & Osorio, 1998). This final point is critical to many tests of ecological and evolutionary hypotheses pertaining to, for example, the efficacy of camouflage (Pessoa *et al.*, 2014; Troscianko *et al.*, 2016), 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), the presence of polymorphism or dichromatism (Schultz & Fincke, 2013; Whiting *et al.*, 2015), or the effect of experimental treatments (Barry *et al.*, 2015). At the heart of these diverse inquiries lies the same question: how different are these colours (or more precisely, these samples of colours) to the animal viewing them? Note that 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; van der Kooi *et al.*, 2016). We thus refer to both as questions of discriminability hereafter, and while the below discussion 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 during 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 as being more generally defined by Weber's law of just noticeable differences $k = \Delta I/I$ — that is, the difference threshold k is a constant determined by the difference between two stimuli ΔI relative to the intensity of the baseline stimulus I. For example, if the difference threshold is $k = 0.1$, then a stimulus I_A will only be perceived as different from I_B in that channel if it is at least 10% greater than I_A (the value of k for any species and receptor usually being determined from behavioural experiments; Vorobyev *et al.*, 2001). Further, the noise e in receptor i is defined by $e_i = k/\sqrt{N}$, where N is the relative abundance of receptor i in the retina. The more abundant a receptor is in the retina relative to the other receptor types, then, the lower the relative noise will be on that channel.

The Weber fraction thus establishes a unit of Just Noticeable Differences (JND's), and distances in colour space can be weighted by photoreceptor noise and expressed in what are essentially units of signal:noise. Values lower than 1 JND represent situations where $\frac{signal}{noise} < 1$ and are predicted to be indistinguishable, while distances close to but greater than 1 JND lie at the very threshold of colour discrimination (conventionally interpreted as the point in which two colours can be distinguished 75% of the time when simultaneously presented against a neutral background; Vorobyev & Osorio, 1998). Values greatly above this threshold (say, above 2 or 4 JND's) are likely so different that they can be told apart with virtually no errors. This provides a useful guide to estimating the similarity of points or groups of points in colour space: the greater the distance between colours, the less alike they are to a given viewer. It follows that if differences between sample A and sample B are, on average, above an established threshold, then we can consider the groups different: sexes dichromatic, subspecies distinct, mimetism imperfect, and so on. This powerful approach allows for a clear link between variation and classification within a perceptual framework, and has for that reason been used in a vast number of studies seeking to answer such questions (Barry *et al.*, 2015; Delhey & Peters, 2008; O'Hanlon *et al.*, 2014; Schultz & Fincke, 2013; White *et al.*, 2017).

This framework, however, raises an important methodological issue: how to adequately compare samples of colours, and estimate if the average distance between them is both statistically and perceptually meaningful (i.e. above-threshold; Endler & Mielke, 2005). Two methods are commonly used. In the first, an "average colour" for each group is derived by averaging their reflectance spectra before modelling the visual system, or by averaging their location in colour space. In either case, this mean quantum catch per-receptor per-group — the centroid for that group in multivariate space — is then used to calculate the colour distance between groups (Fig. 1, bold arrow). There are two issues with this approach. First, since colour distances are perceived in a ratio scale, the centroid obtained from arithmetic means of receptor coordinates is not an appropriate measure of central tendency. Instead, the geometric mean (or the average reflectance of log-transformed spectra, as suggested by Cardoso & Gomes, 2015) must be used. This can be demonstrated by converting perceptual distances into Cartesian coordinates (Pike, 2012), in which case the distance between arithmetic means in this perceptual space matches the distance between geometric means in the untransformed scale. Second, 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).

The second approach calculates the pairwise distances between all points in group A and group B (or between group A and the mean of group B if, for example, group B consists of samples from the background), then using the average of these distances to represent the 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 because it allows the calculation of measures of variation such as the standard error of the distances, and thus a t-test or equivalent can be used to test if these differences are greater than the threshold value. 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 colour space and the direction of the distance vector, reflecting only the magnitude of differences between two points. Therefore, the average linkage reduces to a measure of spread, rather than one of relative position, and will scale with both within- and between-group distances. We can demonstrate this by considering the Euclidean distance in one dimension, in which the distance between points x_1 and x_2 is $\sqrt{(x_1 - x_2)^2}$, which reduces to the absolute difference $|x_1 - x_2|$. The mean absolute difference, in turn, is a measure of dispersion, so if we have two $N = 2$ identical samples $A = B = \{10, 20\}$, the average linkage between these samples would be 5, not zero as expected from comparing two identical samples (Fig. 1).

These methods highlight the fact that appraising hypotheses of discriminability has centred on tests of whether the difference between samples is above a perceptual threshold. However, the ready convenience of such threshold value belies fact that simply comparing mean distances between groups is not sufficient to infer, statistically, whether the samples being compared are different. In order to answer if two groups are different, one must compare the level of between-group variation relative to within-group variation. This is particularly problematic in the case of colours that function as signals, such as those used in social interactions (e.g. Kemp & Rutowski, 2011). For a trait to function as a signal in this context, an observer must be able to tell the difference between signals of "low quality" and "high quality". This means, by definition, that *individuals within a statistical population should be readily distinguishable* — they must be highly variable and colour distances between them should be above the threshold of discrimination (Delhey *et al.*, 2017), otherwise no information can be extracted by a viewer when comparing phenotypes.

This is readily appreciable by considering a hypothetical species that uses colour in mate choice, but is not sexually dichromatic (Fig. 1). In this species colour will be highly variable and, on average, differences among individuals of the same sex will fall above the threshold of discrimination, but there is no *consistent* difference between males and females. Therefore, if a researcher took a sample from this species and calculated the average distance between all pairs individuals, regardless of sex, these differences should be largely greater than 1 JND. However, it also follows that if the researcher took separate samples of males and females from that species, then all pairwise distances (the average link distance) between sexes will be also greater than 1 JND, despite them being sampled from the same (statistical) population.

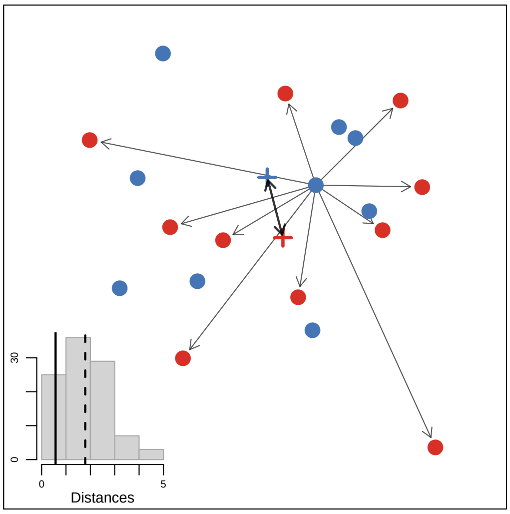

Figure 1: The link distance (i.e. average pairwise distance between groups in a colour space) conflates among- and within-group variation. Here, two samples were drawn from the same distribution. Thin arrows indicate the distances between a random point in the first sample (blue) and all points from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

The issues raised above highlight the fact that identifying differences between groups as lying above or below a perceptual threshold, even with an adequate measure of uncertainty, is still not sufficient to tackle questions of discriminability — it is also essential to consider how the sample are distributed relative to one another in colour space. The importance of considering the distribution of colour points relative to each other when comparing colours has been noted before (Eaton, 2005; Endler & Mielke, 2005). Eaton (2005), for example, noted that within-group variation would influence his conclusions about the extent of avian dichromatism using avian visual models. His solution to the problem was to test for intersexual differences in photon catches separately for each receptor. However, this approach ignores the multivariate nature of visual model data, and may thus inflate Type I errors by failing to account for multiple comparisons and ignoring correlations among receptor catches (which is critical, since any visual system defined by n receptors can be represented in $n-1$ dimensions; Kelber *et al.*, 2003). Further, unless quantum catches are made relative to their sum, this univariate approach also fails to consider that visual models ignore the achromatic dimension of colour and the absolute value of receptor-specific quantal catches (Endler & Mielke, 2005; Vorobyev & Osorio, 1998). That is, a colour that stimulates a bird's four receptors $\{u, s, m, l\}$ by $\{1, 2, 3, 4\}$ should have a distance of zero to a colour $\{10, 20, 30, 40\}$, but univariate analyses that ignore the multivariate structure of colour spaces might conclude otherwise. (Note that this only holds if the Weber fraction is constant, such as under the assumption of a bright illuminant; otherwise the noise is not constant and is indeed influenced by the irradiant intensity. However, even in such case differences would result from signals evoking different noises in the receptor, and thus still making a univariate approach problematic Osorio *et al.*, 2004).

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 the convex hull defined by all its points, and the separation between 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 in this nest parasite interaction. This approach is interesting because it considers the entire 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 measure of their distance, offering no distinction between non-overlapping samples that are near or far from each other in colour space) and (ii) it is unclear how varying degrees of volume overlap should be interpreted biologically (e.g. how biologically meaningful is the difference between 20% or 40% overlap?). It is also particularly sensitive to outliers such that, for example, if two samples largely overlap but one or both include extreme values that "stretch out" their volumes, the overlap between these groups will be underestimated. Likewise, there is no distinction between cases in which there is a small overlap between two samples due to close proximity *versus* two samples that are largely separated but either has extreme values that "reach into" each others' volume. These problems arise because the volume as 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, but 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 doing so 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 approach is powerful due to its multivariate nature and the fact that it calculates a measure of effect size based on the relationship of between- and within-group distances. However, this approach calculates a single effect size statistic that captures differences between samples not only in their means, but also in their dispersion and correlation structure (i.e. shape; Endler & Mielke, 2005). In other words, like many other multivariate distance-based methods, this method is sensitive to confounding heterogeneity among samples when attempting to test for differences in *location* between samples (Anderson & Walsh, 2013; Warton *et al.*, 2012). Further, like the volume overlap, this approach does not consider discrimination thresholds (though, again, it would be straightforward to substitute Euclidean distance for distances in JNDs to obtain receptor-noise limited statistics). Despite its strengths, this method has seen little adoption in the discipline over the last decade, largely due to limitations in implementation and accessibility.

The shortcomings of the methods described above reflect the fundamental fact that the question of discriminability actually represents a test of two hypotheses that are seldom formally distinguished: (i) that the focal samples are statistically distinct, and (ii) that the magnitude of their difference is greater than a psychophysiological threshold of detection. Most approaches will test one, but not both, of these hypotheses through their respective nulls, and more often than not with no estimate of variation, measurement error, or uncertainty in their estimates. Below we use a simulation-based approach to quantify these issues 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, and demonstrate its utility using heuristic examples drawn from the literature.

Methods

Simulation procedures

To test 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 in each 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 groups of colours (N = 50 per group) defined by four variables (*usml* photoreceptors) sampled from a multivariate log-normal distribution (given that quantum catches are non-negative and perceptual distances follow a ratio scale, as defined by the Weber fraction described above). We generated samples according to two different scenarios: first, we simulated groups with 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 focused on testing power and error-rates we sought to consider a wide range of positions in colour space and intra-group variances. We therefore 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). In order 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) from a log-normal distribution with $\mu_{iB} = \mu_{iA} + k_i$ and standard deviation 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), each replicated 200 times for group sample sizes $N = \{10, 20, 50, 100\}$.

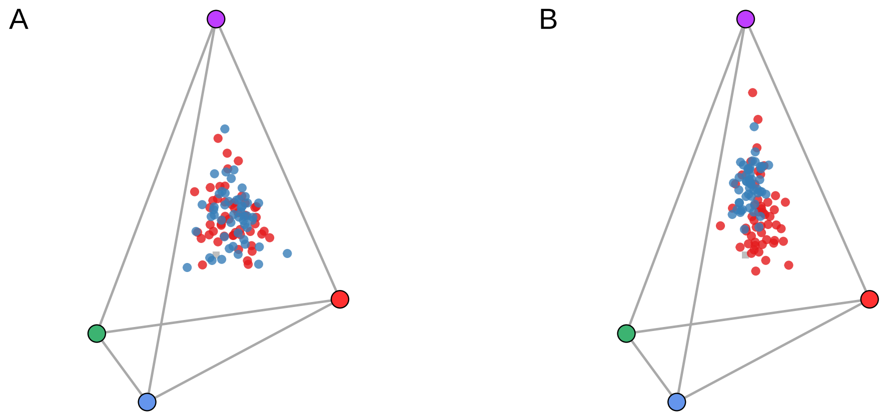

Figure 2: Example simulated data for the two groups (red, blue) in a tetrahedral colourspace. Shown here are data with sample size $N = 50$, and effect size $S = 0$ (left) and $S = 3$ (right).

For the second set of simulations we focused on threshold conditions across a range of within-sample variation, following a similar procedure as described above. 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 (as often assumed for the long-wavelength photoreceptor for birds; Vorobyev *et al.*, 1998), we would need to draw from a distribution that differed in geometric mean quantal catch by $\frac{0.1}{\sqrt{n}}$ (as described above for the relationship between S and k). However, due to variation in relative receptor densities, the Weber fraction will also vary among receptors, even if assuming a constant single receptor noise-to-signal ratio. Therefore, to simplify simulations, we drew group B from a log-normal distribution with $\mu_{iB} = k_i \mu_{iA}$, where $k_i \sim \mathcal{U}(0.88, 1.12)$, resulting in a 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 simulation replicates.

After the two groups were simulated, we used the R package `pavo` (Maia *et al.*, 2013) to calculate the colour distances between each pair of points in colour space. We used the default function parameters (relative receptor densities for $\{u, s, m, l\} = \{1, 2, 2, 4\}$ and Weber fraction for $l = 0.1$). Based on these distances, we calculated the average within-group pairwise distance across both groups, as well as the average between-group pairwise distance (average link distance). Further, we estimated the geometric mean for both groups and the distance between them.

We then used four procedures to statistically test for a difference between the two groups. First, we used a distance-based PERMANOVA (hereafter "distance PERMANOVA") using the `adonis` function in the R package `vegan` (Oksanen *et al.*, 2007). This non-parametric approach uses distances in JND to directly calculate a pseudo-F statistics based on the ratio of among:within distances between groups, and obtains a null distribution by randomizing distances between observations (Anderson, 2005). We recorded if the analysis was significant using 999 permutations for the null distribution, as well as the R^2 (the proportion of dispersion explained by the grouping factor) as an estimate of the effect size of the test. Second, we obtained XYZ Cartesian coordinates based on perceptually-scaled distances Delhey & Peters (2008); Pike (2012), and conducted a MANOVA on these variables, again recording if the analysis was significant (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 that allow for more complex parametrizations and relaxed assumptions. Third, we calculated the volume overlap between the two groups of points (relative to their combined volumes) in a tetrahedral colour space defined by the receptors' relative quantum catches (and thus not considering receptor noise; Stoddard & Prum, 2008). Finally, we repeated the volume overlap for the XYZ Cartesian coordinates based on perceptual distances, thereby generating a colour volume overlap that also 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 groups were sampled from the same population ($S = 0$), even when sample sizes are small (Fig. 3). As expected, the power to detect small effects steadily increased as a function of sample size, with the distance PERMANOVA being slightly more conservative than the Cartesian MANOVA across both sample and effect sizes (Fig. 3,4).

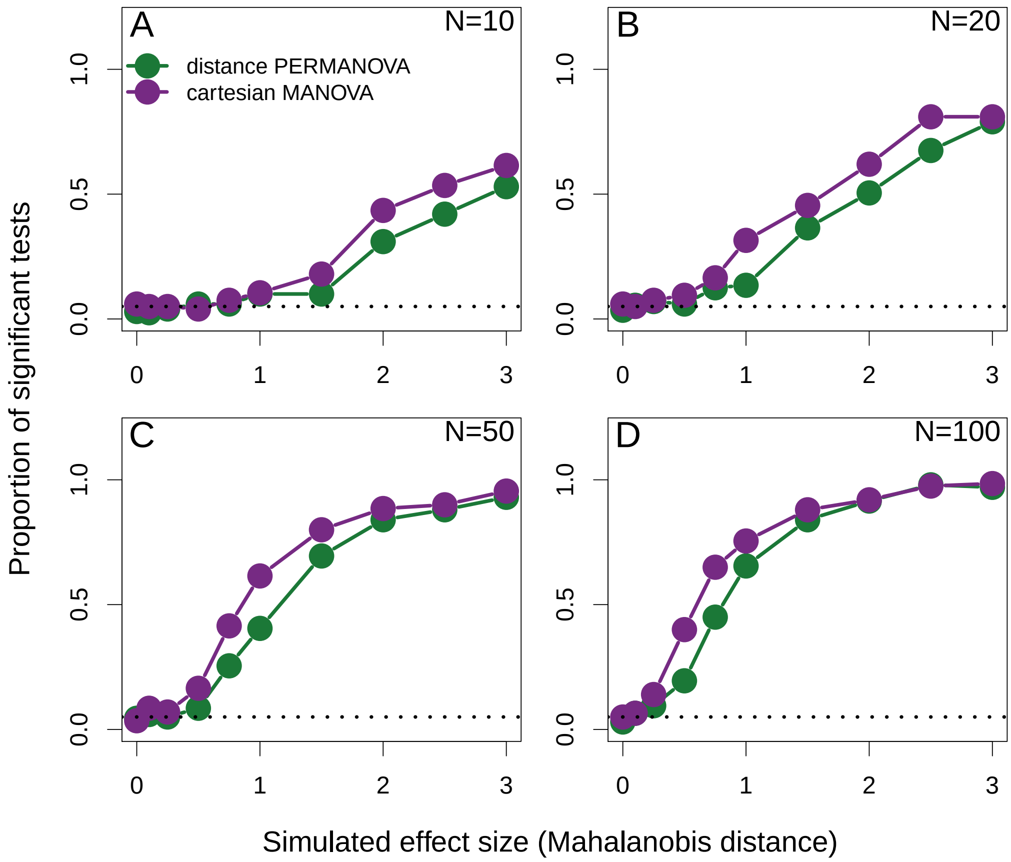

Figure 3: Power and Type I error rate of the distance PERMANOVA (green) and Cartesian MANOVA (purple). Panels show the proportion of simulations yielding significant results for each approach for simulations with varying sample sizes and effect sizes.

As a result, the two approaches showed some disagreement, with between $10 - 15\%$ of the simulations being significant only in one of the two approaches (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 consider a comparison significant when it was not significant under the distance PERMANOVA than vice-versa, at an approximately constant rate regardless of the sample size (Fig.4).

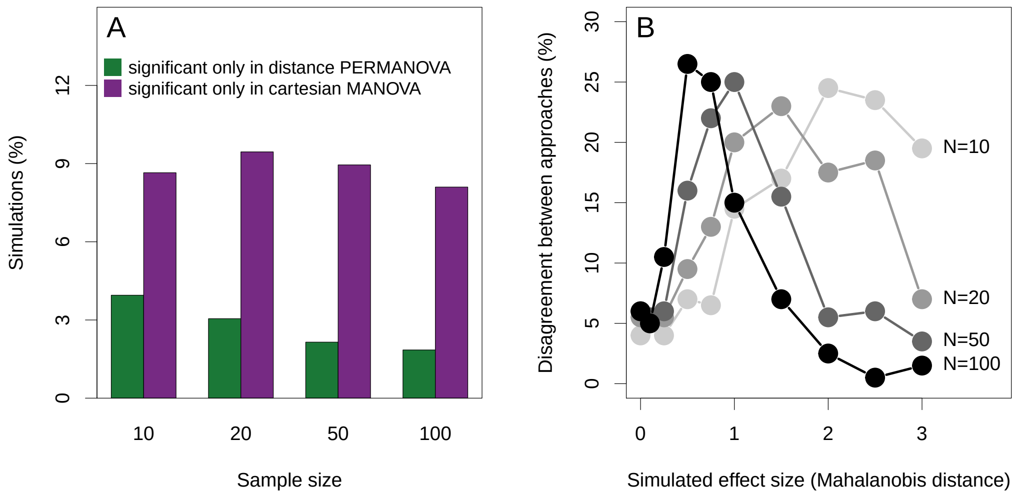

Figure 4: The disagreement between multivariate statistical approaches when testing for separation between samples in colour space.

Focusing on the $N = 50$ simulations, our results show that about half (46.5%) of our simulations produced samples with mean difference greater than $1JND$ (Fig. 5). In these simulations, 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 the $1JND$ mark significance is variable, showing that large within-group variation can lead to non-significant differences between groups despite among-group differences being, on average, above the perceptual threshold. Colour volume overlap also showed a (negative) association with the effect size, but despite being both an estimate of relative position and overlap between groups, no specific threshold for significance is identifiable (for example, both significant and non-significant results are observed for values of overlap between 20 and 60%; Figure 5B).

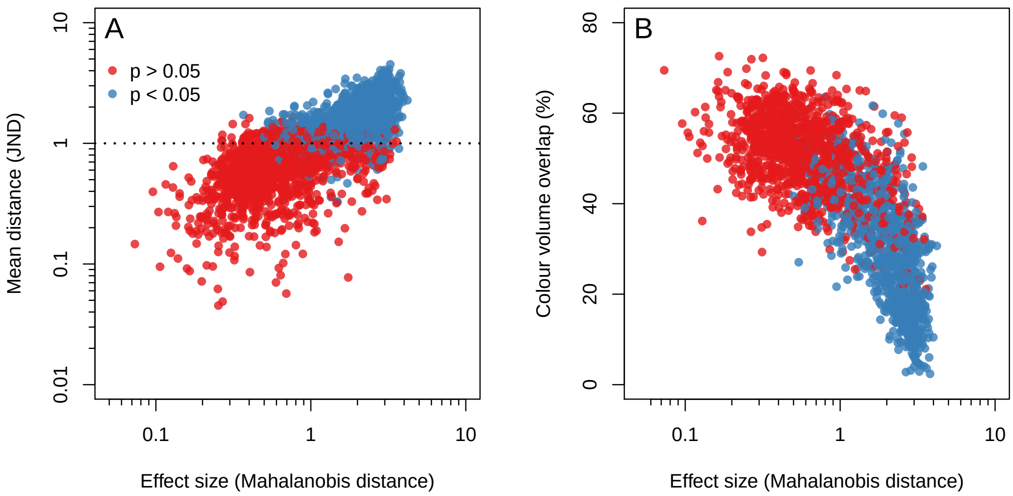

Figure 5: The association between effect size and (A) mean distance and (B) colour volume overlap. Significant distance PERMANOVA results are in blue, whereas non-significant results are in red. Dotted line indicates the threshold of 1JND.

Threshold scenarios

Our second set of simulations evaluated scenarios in which among-group differences were near the threshold of detectability (ca. 1 JND) across a wide range of within-group variation. Since both the distance PERMANOVA and the Cartesian MANOVA produced similar results, we focus on the former due to the convenience of the resulting R^2 statistic in approximating the degree of among-group separation. Our simulations generated a wide range of outcomes as desired, with non-significant and significant tests both above and below the perceptual threshold of 1 JND (Fig. 6). Thus, in contrast with our power simulations above (Fig 5), the significance threshold did not match the perceptual threshold in these simulations. Thus, as in the hypothetical example discussed in the introduction, some of our simulated groups were statistically inseparable despite having mean distances above the perceptual threshold (Fig. 6, dark red points). Likewise, some of our simulations produced scenarios in which the samples were statistically different, but that difference was below the perceptual threshold and therefore biologically undetectable to this observer (Fig. 6, dark blue points). These results highlight the importance of considering both among-group 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 proportionally small relative to among-group differences. However, when measuring within-group difference using the link distance (i.e. the average pairwise distance between all pairs of colours) nearly all simulations—including most significant results—fell below the 1:1 line, indicating that the link distance is a poor approximation of the variance-covariance structure of the data, overestimating it by about $0.5 JND$ (grey line in Fig. 6A: mean distance = mean within-group distance-0.05). We can further 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%).

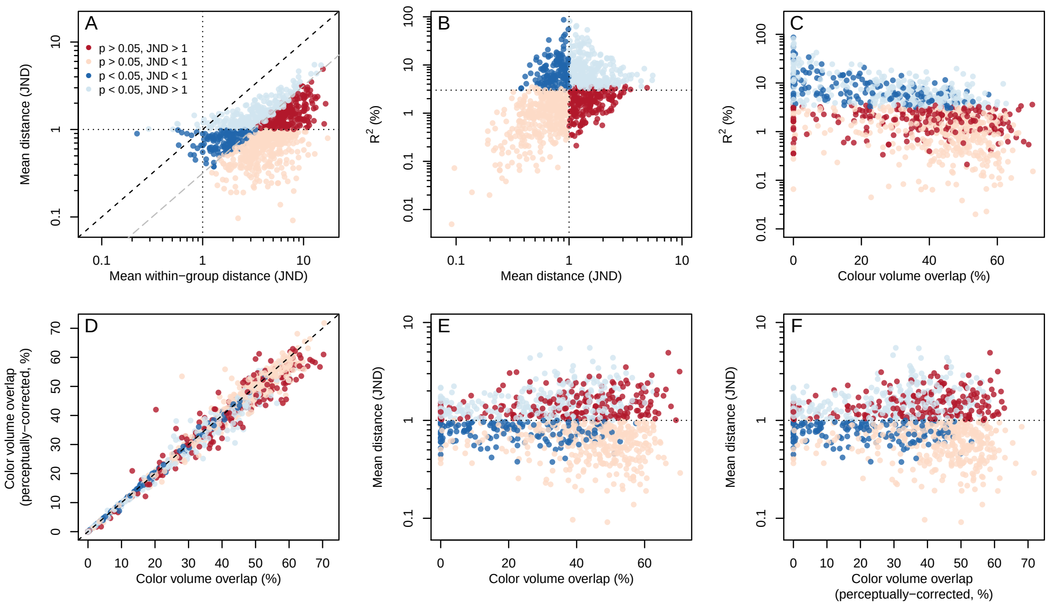

Figure 6: Results from threshold simulation. Red and blue denote non-significant and significant PERMANOVA tests, respectively, and light colours denote when that approach would yield the same inference as comparing mean distances to a threshold of 1JND. Thus, dark blue points indicate a significant 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 the overlap between colour volumes, our results indicate a low overall consistency between these two approaches: for any given level of volume overlap, all possible combinations of results are observed — even when there overlap between samples is zero (Fig. 6C, E-F). In other words, even complete separation in colour volumes can result in non-significant, below-perceptual threshold cases, since samples can be in close proximity in perceptual space (Fig. 6E-F) and have high within-group variance. Likewise, samples can have high overlap but still be statistically and perceptually distinguishable, because their overall distributions are nonetheless discernible in multivariate space. Further, there is no association between volume overlap and mean distance between groups (Fig. 6E,F). Importantly, these results were unaltered by calculating volumes in perceptual colour space, since these are still strongly and positively correlated with their non-perceptual counterparts (Fig.

6D,F).

A flexible method for estimating statistical and perceptual separation

As described conceptually and shown through simulations above, testing for discriminability between two samples of colours actually requires testing two distinct hypotheses: (i) are samples statistically distinct, and (ii) are samples perceptually distinct. We therefore propose a two-step approach to answering the question of discriminability between groups, which explicitly formalizes these hypotheses.

For the first question — are the samples of colours *statistically separate* in colour space? — both a PERMANOVA using perceptual colour distances (Anderson, 2005; Cornuault *et al.*, 2015), or a MANOVA using perceptually-calibrated Cartesian coordinates (Delhey & Peters, 2008; Pike, 2012) are well suited to the task (demonstrated above). Both approaches exclude achromatic variation, properly account for the multivariate nature of visual model data, and perform well when facing heterogeneity and outliers. There are also minimal difference in results between the two (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 has been shown to be 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 relatively common among colour space 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 as part of the first question will not generally be applicable here, since measures of effect size desirably account for the ratio of among:within variation. We therefore suggest this be tested independently by estimating the distance in colour space between group geometric means rather than by calculating the average distance of all pairwise comparisons or volume-overlap based metrics, which fail to accurately estimate group separation (Fig. 6). However, this approach still has the limitation of generating a single measure of distance for each pair of groups being compared, with no measure of uncertainty. We thus suggest a bootstrap procedure in which new samples for each group (of the same size as the original groups) are produced through a re-sampling procedure (with replacement) of individuals from that group, from which geometric means and the distance between them are calculated. This procedure generates a distribution of mean distances, from which the confidence interval for the observed distance 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 estimate their degree of dissimilarity.

Empirical examples

We present two brief examples of this two-step approach to the analysis of visual model data below, centred on questions of near-threshold discrimination drawn from the literature. As above, we used the `R` package `pavo` for visual modelling, and the `adonis` function in the `R` package `vegan` for PERMANOVAs.

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 of available defences (Barry *et al.*, 2015; Grether *et al.*, 2004; Johnstone, 1995). The nature of intraspecific variation in colour traits can thus act as a guide to their putative function, since selection may act differentially on signals used in different contexts. Aposematic signals, for example, may be relatively invariable within species, by virtue of their reliance on the formation of learning rules by predators (Endler, 1992; Guilford, 1990). Traits subject to strong sexual selection in one of the sexes, in contrast, are often characterised by dimorphism, in which one sex (typically males) expresses a conspicuous colour pattern that is greatly reduced or absent in the other (Bell & Zamudio, 2012; Kemp & Rutowski, 2011).

Dragon lizards (Agamidae) are well known for variable colouration that is 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 sexual dichromatism among lizard body regions to test which colour patches, if any, are sexually dimorphic from the perspective of conspecifics.

Following standard calculations for the log-linear receptor-noise model, we estimated cone quantum catch as the log-transformed integrated product of stimulus reflectance, ambient illumination, and photoreceptor absorbance across the 300-700 nm waveband (Vorobyev *et al.*, 1998). 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 photoreceptor signal-to-noise ratio that yielded 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 PERMANOVAs with 999 permutations.

We found a statistical difference between male and female throats (PERMANOVA: $F_{1,58} = 14.84$, $P < 0.01$) and labials (PERMANOVA: $F_{1,57} = 13.96$, $P < 0.01$) as modelled in agamid colourspace (Fig. 7a, b), but not for tongues (PERMANOVA: $F_{1,58} = 1.63$, $P = 0.22$) or mouth-roofs (PERMANOVA: $F_{1,55} = 0.52$, $P = 0.50$; Fig. 7C,D). However, subsequent bootstrap-based analysis of group separation

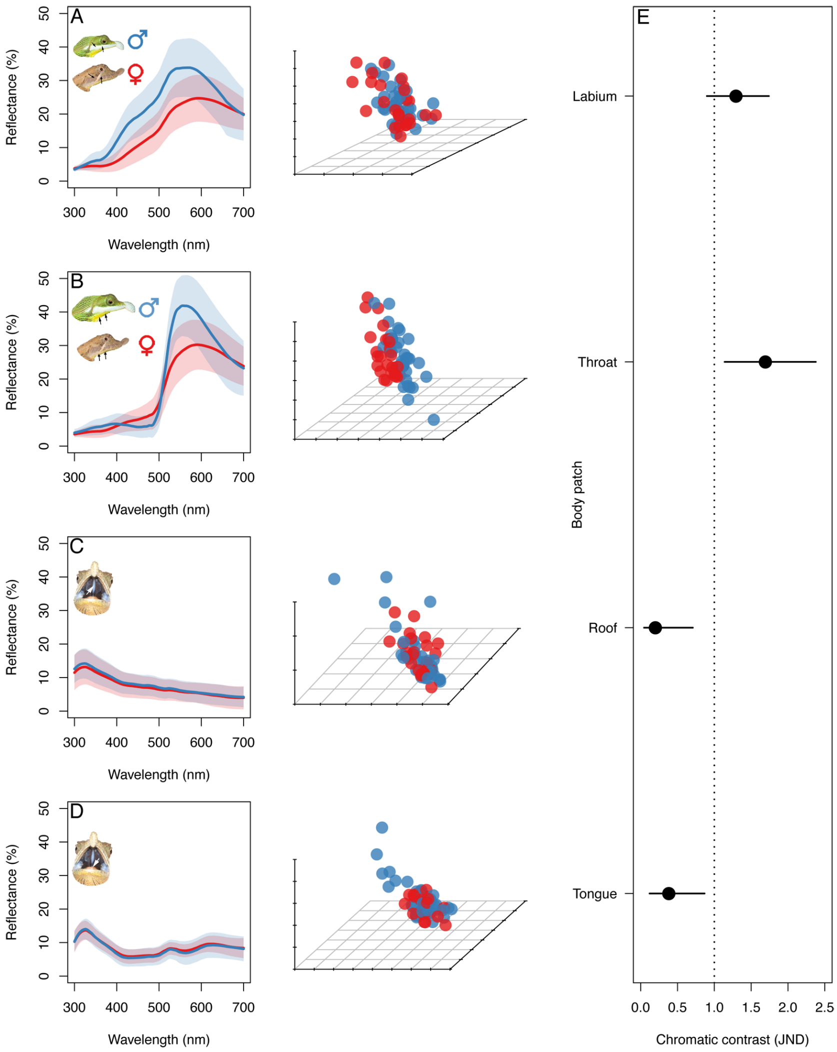

Figure 7: The mean (± SD) spectral reflectance of female (red) and male (black) (A) labial, (B) throat, (C) mouth-roof, and (D) tongue (left panels), and their colourspace distribution according in a tetrachromatic model of agamid vision (middle panels). Inset images indicate approximate sampling regions. 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.

suggested that intersexual differences in labial colour are likely imperceptible to conspecifics, and throat colour differences only barely so (Fig. 7; note that we present mean JND differences between all body parts for illustrative purposes, and it would generally be unnecessary to pursue this step for tests that were non-significant in the first step). Our results therefore suggest the absence of dichromatism in most measured body regions from the perspective of conspecifics despite statistically significant differences in their colours, with only a subtle (predicted) perceivable difference on the labials of males and females. Thus these results do not implicate sexual selection as a strong driver of intersexual colour differences in these body regions of *C. ornatus*.

Floral mimicry in the spiny spider *Gasteracantha fornicata*

Biases in sensory perception offer opportunities for the evolution of deception (Endler & Basolo, 1998). This is often showcased in predator-prey interactions, wherein predators induce maladaptive responses in prey by using signals that exploit innate preferences for certain visual cues, or learned preferences for otherwise rewarding stimuli (e.g. Heiling *et al.*, 2003; OHanlon *et al.*, 2014). Many sit-and-wait predators, such as orb-web spiders, use conspicuous colouration to visually lure prey (reviewed in White & Kemp, 2015). While the attractant nature of these signals is well documented (e.g. Chuang *et al.*, 2008; Tso *et al.*, 2007, 2002), the ultimate basis of their effectiveness — that is, the nature of the sensory or perceptual pathways being exploited in prey — remains unclear. A long-standing hypothesis is that lures have evolved to mimic flowers (Chiao *et al.*, 2009; Tso *et al.*, 2004), in which case theory predicts that the signals of mimics (lures) and sympatric models (flowers) should be largely indistinguishable to their shared receivers (Christy, 1995; Endler & Basolo, 1998).

We tested this hypothesis using *Gasteracantha fornicata*, a conspicuously coloured orb-web spider found in tropical and sub-tropical forests of Australia. Females of this species are stably polymorphic and exhibit either 'white' or 'yellow' (UV-) bands against a black outline (Fig. 8 inset), which serve to attract insect prey (Hauber, 2002; Kemp *et al.*, 2013). We used a subset of data from several recent studies Dalrymple *et al.* (2015); White *et al.* (2017); White & Kemp (2016), which included reflectance spectra from the coloured bands of 33 yellow and 29 white *G. fornicata*, and the flowers of 36 sympatric angiosperm species. We tested whether spiders were discretely polymorphic from the perspective of a representative taxon

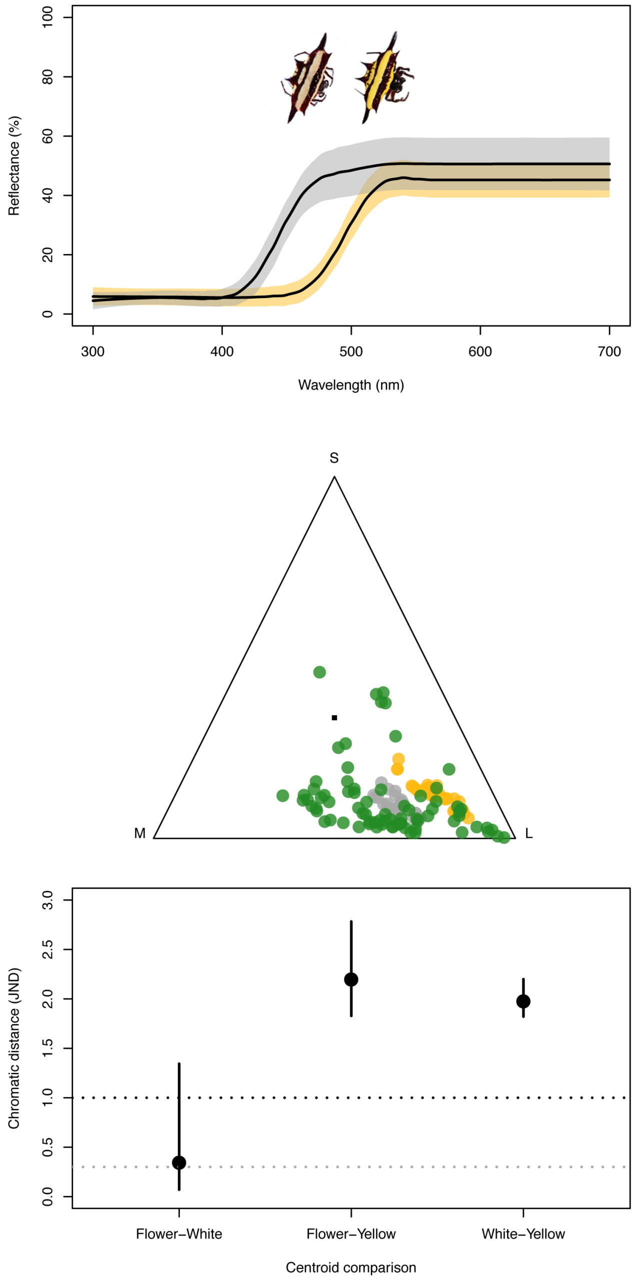

Figure 8: The mean (± SD) spectral reflectance of both morphs of the spiny spider *Gasteracantha fornicata* (top), and the location of spectra in a trichromatic colourspace representing Honeybee (*Apis meliffera* vision (middle), where yellow and grey points denote 'yellow' and 'white' spider morphs, respectively, and green points are sympatric flowers. The bootstrapped 95 % C.I. for mean distances between groups in honeybee colour space (bottom). Dotted lines indicate thresholds of 1JND (black) and 0.3JND (grey).

of insect prey (as given their conspicuous differences to human viewers; Fig. 8), and whether spiders were indistinguishable from sympatric flowers, as predicted by a floral-mimicry hypothesis (Christy, 1995; White & Kemp, 2015).

To estimate discriminability according to the receptor-noise limited model, we used the visual phenotype of the honeybee *Apis melifera*, with $\lambda_{max} = 340, 440, 536$ nm based on a vitamin A1 template (Govardovskii *et al.*, 2000). We estimated photoreceptor noise using a Weber fraction for the long wavelength receptor of 0.13, and a ratio of 1:0.5:4.4 for the relative density of "ultraviolet", "blue", and "green" photoreceptors. To test our hypotheses, we specified a PERMANOVA with *a priori* two statistical contrasts: in the first, we considered yellow-versus-white spider morphs (to test for polymorphism), and in the second, each morph-versus-flowers (to test for mimicry).

We found spider morphs to be statistically distinct from one another in honeybee colour space (PERMANOVA: $F_{1,60} = 32.13$, $P < 0.01$), but only the yellow morph was statistically different from sympatric flowers (PERMANOVA: white vs. flowers: $F_{1,134} = 1.61$, $P = 0.21$; yellow vs. flowers: $F_{1,134} = 9.30$, $P < 0.01$). Our bootstrapped group distances predict that the statistically significant differences are also perceptually distinct to a honeybee observer (Fig. 8). At a glance, these results offer relatively weak support for the hypotheses of polymorphism and floral mimicry. As in all such analyses it is of course necessary to consider the assumptions of the model used. This includes consideration of the relevant "threshold" value for focal species, which—even when based on an accurately parametrised model (e.g. using precise Weber fractions)—may not strongly predict an animal's realised abilities (Dyer, 2012; Dyer & Neumeyer, 2005). Honeybees, for example, are capable of discriminating between simultaneously presented colour-stimuli that differ by only 0.3 JND's under laboratory conditions (Dyer & Neumeyer, 2005). In this case, our results would be largely unaffected, because the statistical test for the difference between white morphs and flowers is non-significant, and the confidence interval of the white-vs-flower comparison also encompasses the laboratory-ideal value of 0.3 JND's. Furthr interpreting the biological implications of such results demands a nuanced consideration of the assumptions of the underlying model, as well as our broader understanding of the ecology of the focal question. For example, future analyses may also consider the perspectives of different viewers, or the distance estimates of alternate visual models (e.g. Chittka, 1992), all of which may be incorporated in our suggested framework.

Discussion

Visual models offer a powerful tool for quantifying the subjective perception of colour which—as the ultimate canvas for colour-signal evolution—affords us 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 samples being compared as well as the limits to perception that observers experience. In this study, we highlight the importance of partitioning these two facets. We show that contemporary methods typically consider only one of these aspects, with undesirable consequences, and propose a flexible, robust alternative that explicitly addresses both.

The use of relatively simpler visual models that do not consider the role of receptor noise when defining colour spaces (and colour differences) is often justified on the basis of relaxing assumptions about the role noise plays in colour perception; a phenomenon that requires intricate empirical work to estimate these parameters with precision and identify the level of naturally occurring variation (Kelber *et al.*, 2017; Olsson *et al.*, 2015; Vorobyev & Osorio, 1998). However, we contend that these simplifying models often make very strong implicit assumptions, which are not necessarily supported by the empirical evidence: namely 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 of detection is essential for tests of discriminability.

Our simulations show that both the distance PERMANOVA and the Cartesian MANOVA perform similarly well in statistically differentiating colours in a perceptual space. As expected, the distribution-free non-parametric approach showed slightly inferior power as a consequence of the relaxed assumptions about underlying distributions. Both approaches are very flexible and can accommodate complex models, such as multiple predictors, interactions and hierarchical designs. Several studies have pointed out that distance-based methods perform poorly when the experimental design is unbalanced or when there are mean-variance relationships or other sources of 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), by reducing the number of variables to a distance alone.

Multivariate generalizations of generalized linear models might offer a flexible alternative, though they are still subject to assumptions of multivariate normality and equality of covariance matrices (though the latter can be relaxed when the largest sample size group has greater absolute values in its covariance matrix; Delhey & Peters, 2008; Tabachnick *et al.*, 2001). These models can also be easily extended to include various error and model structures, such as hierarchical and phylogenetic models (Hadfield & Nakagawa, 2010; O'Hara & Kotze, 2010; Warton *et al.*, 2012), and multi-response models can also relax the assumptions of heteroscedasticity by estimating the variance-covariance of response variables (Hadfield, 2010). When using a Bayesian approach, the mean distance bootstrap can also be substituted by estimating distance credible intervals from the posterior distribution of perceptually-corrected Cartesian coordinate estimated centroids, though this will also be influenced by the priors adopted. This approach also allows for the straightforward inclusion of luminance (achromatic) differences as another axis of variation in multivariate analyses (Pike, 2012), and can be parametrized to reflect specific aspects of colour perception and discriminability. For example, when comparing males and females of a species of bird with multiple coloured patches, we may be interested in whether there is total plumage differences between the sexes or, on the other hand, if there is at least a plumage patch that distinguishes the sexes. By parametrizing a hierarchical model with body patch as a random or fixed effect, one can model precisely what is meant when asking if the species is dichromatic.

Delhey & Peters (2008) have recently advocated a similar approach to answer the first question in our two-step approach, by suggesting the application of a Principal Component Analysis (PCA) to the perceptually-corrected Cartesian coordinates as an intermediate step before a MANOVA. However, if all the principal components are used in the multivariate analysis, results will be numerically identical to simply using the XYZ coordinates directly. Further, Since it is often tempting to discard PC axes of low variance in downstream analyses, which could be problematic given the roll that residual variance may play in among-group differentiation, we recommend using the Cartesian coordinates directly. Still, using PCA's that preserve colour distances may be particularly useful when investigating differences in the orientation of axis of variation, and may be more readily interpretable when tested against continuous variables (Delhey & Peters, 2008). Of course while we have focused on tests of differences in the multivariate 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.

These approaches will only test the degree of separation between groups in colour space, and so it is still necessary to provide an estimate of the magnitude of that separation. The bootstrap approach we present here provides a simple solution, by adding an easy to interpret measure of accuracy to the mean distance estimate. It is essential, however, to parametrize the underlying visual model appropriately. The Weber fraction chosen for the receptor noise will strongly affect perceptual distances (Bitton *et al.*, 2017) since it directly scales with the JND unit. Further, even when adequate values of the Weber fraction are used, it is important to realize that the unit JND value usually reflects psychophysiological limits under extremely controlled conditions (Kelber *et al.*, 2003; Olsson *et al.*, 2015), and that a more conservative estimate of two, four or even greater 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 assumptions implicit in questions of discriminability. By rendering these assumptions explicit, our two-step approach offers a simple, 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 inta- and interspecific interactions, and provide an efficient analytical framework for the study of colour in nature.

Implementation

All analyses conducted here can be found in the project's GitHub page, and will be implemented in a release of the `R` package `pavo` to accompany the publication of this manuscript. For now, they can be found in the GitHub page, under the `/R/` folder. The function `bootcentroiddS` conducts the boostrap for the calculation of confidence intervals for mean distances, and the function `jnd2xyz` converts chromatic distances in JNDs to perceptually-corrected Cartesian coordinates. These functions require the bleeding edge version of `pavo` to work, which can be found on GitHub: `https://github.com/rmaia/pavo/`.

Acknowledgments

We would like to thank Ruchira Somaweera for permission to use 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. Both authors contributed critically to the drafts and gave final approval for publication.

Data accessibility

Leaf-nosed lizard colour data from Whiting *et al.* (2015) is openly available at http://dx.doi.org/10.6084/m9.figshare.1452908. Floral and spiny spider data from Dalrymple *et al.* (2015) and White *et al.* (2017) is available from http://dx.doi.org/10.6084/m9.figshare.1517656.v1 Simulation data is available from the project GitHub page at https://github.com/rmaia/msdichromatism/.

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.

Chiao, C.C., Wu, W.Y., Chen, S.H. & Yang, E.C. (2009) Visualization of the spatial and spectral signals of orb-weaving spiders, nephila pilipes, through the eyes of a honeybee. *Journal of Experimental Biology*, **212**(14), 2269–2278.

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.

Christy, J.H. (1995) Mimicry, mate choice, and the sensory trap hypothesis. *American Naturalist*, 171–181.

Chuang, C.Y., Yang, E.C. & Tso, I.M. (2008) Deceptive color signaling in the night: a nocturnal predator attracts prey with visual lures. *Behavioral Ecology*, **19**(2), 237–244.

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).

Dalrymple, R.L., Kemp, D.J., Flores-Moreno, H., Laffan, S.W., White, T.E., Hemmings, F.A., Tindall, M.L. & Moles, A.T. (2015) Birds, butterflies and flowers in the tropics are not more colourful than those at higher latitudes. *Global Ecology and Biogeography*, **24**(12), 1424–1432.

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).

Dyer, A.G. (2012) The mysterious cognitive abilities of bees: why models of visual processing need to consider experience and individual differences in animal performance. *Journal of Experimental Biology*, **215**(3), 387–395.

Dyer, A.G. & Neumeyer, C. (2005) Simultaneous and successive colour discrimination in the honeybee (apis mellifera). *Journal of Comparative Physiology A*, **191**(6), 547–557.

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.

Endler, J.A. (1992) Signals, signal conditions, and the direction of evolution. *American Naturalist*, **139**, S125–S153.

Endler, J.A. & Basolo, A.L. (1998) Sensory ecology, receiver biases and sexual selection. *Trends in Ecology & Evolution*, **13**(10), 415–420.

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.

Guilford, T. (1990) The evolution of aposematism. *Insect defenses: adaptive mechanisms and strategies of prey and predators*, 23–61.

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.

Hauber, M.E. (2002) Conspicuous colouration attracts prey to a stationary predator. *Ecological Entomology*, **27**(6), 686–691.

Heiling, A.M., Herberstein, M.E. & Chittka, L. (2003) Pollinator attraction: crab-spiders manipulate flower signals. *Nature*, **421**(6921), 334–334.

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., Holmes, C., Congdon, B.C. & Edwards, W. (2013) Color polymorphism in spiny spiders (gasteracantha fornicata): testing the adaptive significance of a geographically clinal lure. *Ethology*, **119**(12), 1126–1137.

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).

OHanlon, J.C., Holwell, G.I. & Herberstein, M.E. (2014) Pollinator deception in the orchid mantis. *The American Naturalist*, **183**(1), 126–132.

O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**(2), 118–122.

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.

Tabachnick, B.G., Fidell, L.S. & Osterlind, S.J. (2001) *Using multivariate statistics*. Allyn and Bacon Boston.

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

Tso, I.M., Huang, J.P. & Liao, C.P. (2007) Nocturnal hunting of a brightly coloured sit-and-wait predator. *Animal Behaviour*, **74**(4), 787–793.

Tso, I.M., Lin, C.W. & Yang, E.C. (2004) Colourful orb-weaving spiders, nephila pilipes, through a bee's eyes. *Journal of Experimental Biology*, **207**(15), 2631–2637.

Tso, I.M., Tai, P.L., Ku, T.H., Kuo, C.H. & Yang, E.C. (2002) Colour-associated foraging success and population genetic structure in a sit-and-wait predator nephila maculata (araneae: Tetragnathidae). *Animal Behaviour*, **63**(1), 175–182.

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.

White, T.E. & Kemp, D.J. (2015) Technicolour deceit: a sensory basis for the study of colour-based lures. *Animal Behaviour*, **105**, 231–243.

White, T.E. & Kemp, D.J. (2016) Color polymorphic lures target different visual channels in prey. *Evolution*.

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.