Behavioral Ecology Advance Access published June 3, 2015

Behavioral
Ecology

The official journal of the
ISBE
International Society for Behavioral Ecology

Behavioral Ecology (2015), 00(00), 1–10. doi:10.1093/beheco/arv072

Original Article

An empirical test of 2-dimensional signal detection theory applied to Batesian mimicry

David W. Kikuchi, Gaurav Malick, Richard J. Webster, Emilee Whissell, and Thomas N. Sherratt
Department of Biology, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada

Received 17 January 2015; revised 28 April 2015; accepted 4 May 2015.

Signal detection theory (SDT) has been invoked to help explain why imperfect mimics of particularly unprofitable or abundant models might experience no further selection to improve their mimicry. However, most tests of SDT have focused on single dimensions of mimetic phenotypes, or used multivariate techniques to compress many dimensions of phenotype into a single scale. Here, we explicitly tested SDT in both one and two dimensions by asking human subjects to discriminate computer-generated mimics and models that varied continuously in both size and/or color. We arrived at two major conclusions. First, although subjects can use prey size or color to help discriminate profitable and unprofitable prey that vary in only one dimension, responses of subjects to prey that vary in two dimensions are poorly represented by multidimensional SDT. Second, because different individuals within groups may use different strategies, the behavior of groups is often better fit by more complex models. In general, humans give more weight to color when making discriminations than is optimal. This bias may indicate that they believe that color has higher relative validity than size. More studies on the behavior of natural predators when foraging on multidimensional prey are urgently needed.

Key words: **adaptation, animal communication, categorization, discrimination, overshadowing.**

INTRODUCTION

In Batesian mimicry, an undefended species (the mimic) evolves a resemblance to a defended one (the model), gaining protection from predators (Bates 1862; Ruxton et al. 2004). Batesian mimicry is a quintessential example of the power of natural selection to produce adaptation, but the extent of this mimicry is often imperfect (Kikuchi and Pfennig 2013). Indeed, there are numerous instances in which mimics possess some degree of similarity to their models but nonetheless can readily be distinguished (e.g., Dittrich et al. 1993; Kikuchi and Pfennig 2010a; Iserbyt et al. 2011). The maintenance of imperfect mimicry poses a fundamental problem because it challenges our intuition about the adaptive process—namely, that natural selection should continually refine an adaptation.

Many hypotheses have been proposed to explain imperfect mimicry, and among the most successful has been the relaxed selection hypothesis (Bates 1862; Duncan and Sheppard 1965; Sherratt 2002; Lynn 2005; Penney et al. 2012). Modern versions of the relaxed selection hypothesis of imperfect mimicry are typically framed in terms of signal detection theory (SDT; Swets et al. 1961; Wiley 1994; Lynn and Barrett 2014), in which receivers seek to discriminate between profitable and unprofitable stimuli using acceptance/rejection criteria that maximize their net payoff (Oaten et al. 1975; Getty 1985; Johnstone 2002; Sherratt 2002; Holen and Johnstone 2004). The SDT approach is utilitarian in that it incorporates not just the likelihood that a perceived stimulus is a profitable mimic ("target") or unprofitable model ("distractor"), but also the benefits of correct decisions (hits and correct rejections), along with the costs of false positives (false alarms) and false negatives (misses)—see Figure 1.When the unprofitable model is particularly costly to attack compared with the benefits of attacking a mimic (Goodale and Sneddon 1977; Penney et al. 2012), or particularly common in relation to the mimic (Lindström et al. 1997; Harper and Pfennig 2007; Kikuchi and Pfennig 2010b; Iserbyt et al. 2011), then there may be no further selection to improve the extent of mimicry once mimics achieve an approximate resemblance to their models, because even a vague resemblance to the models is sufficient to protect mimics from attack.

Multidimensional signals are ubiquitous in communication (Rowe 1999; Hebets and Papaj 2005), including mimicry, yet SDT models have tended to represent the perceived stimuli of targets and distractors on a 1-dimensional scale, generally thought of as "sensory magnitude." Indeed, even stimuli well known to vary along multiple dimensions are frequently collapsed onto a single appearance dimension. For example, mammogram features such as shape, size, and darkness can effectively be collapsed into a single appearance dimension using multivariate techniques (Dorsi and Swets 1995). Likewise, multivariate techniques have been employed to assess the similarity between mimics and models based on a range of phenotypic properties (e.g., Iserbyt et al. 2011; Penney et al. 2012).

Address correspondence to D.W. Kikuchi. E-mail: dwkikuchi@gmail.com

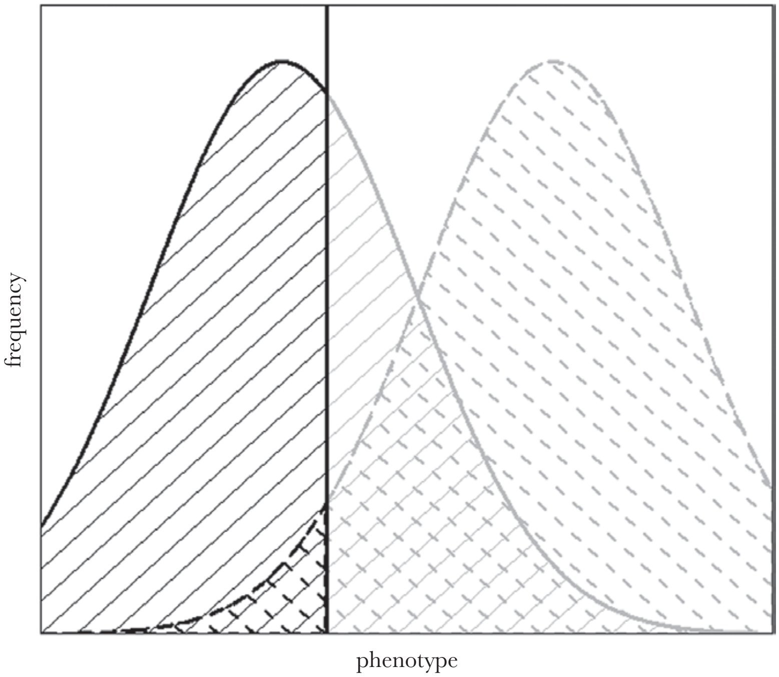

Figure 1
All signal detection problems are based on a set of probability distributions of perceived stimuli, generated by targets and distractors. These probability distributions can take any form but the most commonly assumed distribution is the Gaussian (normal), parameterized by the mean (μ), and SD (σ). When the distributions of perceived stimuli from a single target and single distractor are each unimodal (e.g., normally distributed), then there is an optimal threshold (also called a decision boundary) at which the receiver should switch from responding as if the stimulus was generated by a target, to as if it were generated by a distractor. This process is like choosing a significance threshold in statistics, except that instead of using an arbitrary number like 0.05, the decision boundary is calculated to maximize the net payoff to the decision maker. In this hypothetical scenario, targets (mimics) are represented by the solid line and distractors (models) are represented by the dashed line. The fills indicate correct attacks on mimics (black thin lines), incorrect attacks on models (black dashes), correct rejections of models (gray dashes), and incorrect rejections of mimics (gray thin lines). Cost = 3, benefit = 1, and number of mimics is equal to the number of models.

Extending SDT to multiple dimensions is a natural alternative to reducing the phenotypic dimension of appearance to a single sensory magnitude. Indeed, when multidimensional traits (such as color and size) are uncorrelated across phenotypes and perceived in an uncorrelated way (such that red objects do not appear any bigger or smaller than blue objects, say), then SDT can divide the perceptual plane into response regions of accept (attack) or reject just as it does in the 1-dimensional case (Ashby and Townsend 1986; Ashby and Soto 2015). When the perceived or actual target and distractor traits are uncorrelated and normal in their distributions, the threshold is a straight line (Figure 2).

Although there has been a great deal of work on 1-dimensional SDT and some experimental evaluations (e.g., Duncan and Sheppard 1963; McGuire et al. 2006; Leonard et al. 2011), there has been very little work in ecology to evaluate the success of SDT applied directly to multidimensional problems. Indeed, despite the prevalence of phenotypic variation in multiple dimensions including size, color, and pattern, it is quite possible that classical SDT fails in multiple dimensions. For example, the framework may assume too much from the decision maker's cognitive abilities in expecting them to weight different phenotypic dimensions to different extents. Only through experimental evaluation of these theoretical models will we fully understand the general limitations of multidimensional SDT and in turn how predators might be expected to respond to mimicry in multiple dimensions.

Here, we conducted an experiment with human subjects as surrogate predators who were asked to discriminate between models and mimics that varied in either one or two independent dimensions (size and/or color). Subjects played a computer game where they received points for correctly attacking mimics and lost points for attacking models. We had two overarching goals: 1) to determine if individual subjects behaved according to the expectations of multidimensional SDT, and 2) elucidate the nature of selection for mimicry by analyzing the overall attack rates on mimetic populations. SDT assumes that subjects perceive and weight all dimensions appropriately in making decisions. To test this fundamental assumption, we varied the degree to which each dimension could be used to make a correct choice. To address our second goal, we measured the attack rates experienced by a population of mimics when they were exposed to predation by groups of subjects. To evaluate the robustness of our findings, we also measured the effect of varying the cost:benefit ratio of attacking models versus mimics on the overall attack rates on mimics.

METHODS

Experimentation

Our computer program generated digital prey (models and their mimics) and displayed them for human subjects to attack. Attacking a model penalized the subject in numerical points, while attacking a mimic gave him or her a reward (cf., McGuire et al. 2006). Prey items were squares that could vary in two dimensions: size (number of tiles ["pixels"] along each of the sides, with 24 pixels appearing as 1.3 cm on the screen) and the proportion of blue pixels (others were yellow). In the color dimension, prey were bilaterally symmetrical with a 50% probability of having a vertical versus horizontal axis of symmetry. In each dimension, the phenotypes of individual models and mimics were drawn from normal distributions with different means but equal variances (a common simplifying assumption in SDT). At the start of each trial, subjects were simultaneously shown 28 randomly selected exemplars of costly models and 28 beneficial mimics that they could study for as long as they wanted until proceeding (Figure 3). They were also informed of the relative costs and benefits of attacking models and mimics, and instructed to maximize their scores. Subjects were then given 5 min to forage on the digital prey, which were presented one at a time in a random order. Each prey was produced by a random draw from the specified size and color normal distributions. For each prey that was presented, subjects could choose to attack it or skip it. There was no time limit imposed in making any individual decision to attack or reject, but we recorded the time (in milliseconds) to attack each prey item and the time until a new screen (with new prey) was requested. After attacking a prey item, subjects' running point total was updated (which reinforced the cost:benefit of attacking prey). To enhance the motivation of our subjects and provide instant feedback on the profitability of prey items they had just attacked, attacks on models were accompanied by an unpleasant electric shock sound, whereas attacks on mimics were accompanied by a pleasant cash-register sound. Subjects' total score was constantly displayed in the corner of the computer monitor. Each subject participated in only one trial.

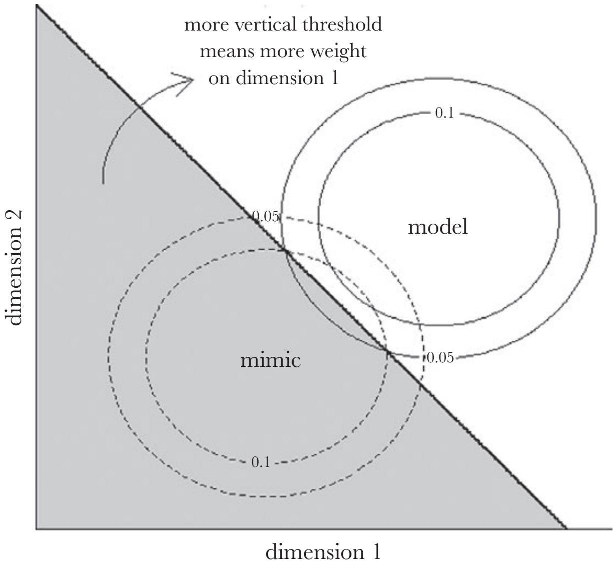

Figure 2
When there are two multivariate normal distributions, with each phenotypic dimension perceived independently (i.e., the two perceptual distributions have equal variance-covariance matrices), the decision boundary for optimally separating targets and distractors is a straight line through phenotypic space. On one side of the decision boundary (thick black line), all targets should be attacked (gray area), and on the other, all should be avoided (white area). Comparing slope and intercept of the optimal decision boundary relative to actual predators' responses is one way to assess how well SDT predicts behavior. For the hypothetical scenario depicted here, the densities of targets (mimics) are represented by the dashed contours and the densities of distractors (models) are represented by the solid contours. If a predator used dimension 1 more than was optimal, its decision threshold would be more vertical than the optimal one. Here, cost = 2, benefit = 1, and number of mimics is equal to the number of models. If relative abundances or costs were to change, we would expect a change in the *intercept* of the decision boundary rather than the slope, so in this case it would remain perpendicular to the line between the model and mimic mean phenotypes.

We had 9 treatment conditions under which subjects were tested (see Table 1 and Supplementary Figure 1). In the first 4, only one dimension (color or size) was informative, and the other dimension did not vary. In treatments 5–9, both color and size were informative (albeit to different extents), that is, subjects' performance on 2-dimensional SDT tasks was tested. Our treatments are summarized in Table 1. For treatments 1–7, three levels of cost:benefit ratio were tested, but for treatments 8 and 9, only one cost:benefit ratio was tested. We refer to the collections of subjects randomly assigned to each treatment at each level of cost:benefit as "groups" (which we use for statistical purposes, rather than to imply group foraging). There were seven subjects for each group, including three groups each under treatments 1–7, and one group each under treatments 8 and 9 (23 groups and 161 subjects in total). Subjects came from the population of Carleton University's campus, where we asked individuals to participate using a portable computer terminal.

One-dimensional SDT

It is possible that humans are simply less able to use one dimension than another to make discriminations, so we conducted trials in which mimics and models differed on average in a single appearance dimension. Both prey phenotypes resembled each other

Table 1
Parameters of experimental treatments used in this study. Treatments 1–4 are for 1-dimensional discrimination tasks

Treatment	Model % blue	Mimic % blue	σ color	Model size	Mimic size	σ size	Description
1	70	40	0.15	24	24	0	Color only (easy to distinguish)
2	60	50	0.15	24	24	0	Color only (hard to distinguish)
3	50	50	0	31.2	24	4	Size only (hard to distinguish)
4	50	50	0	36	24	4	Size only (easy to distinguish)
5	70	40	0.15	31.2	24	4	Color and size moderate
6	60	50	0.15	36	24	4	Color hard; size easy
7[a]	40	70	0.15	24	31.2	4	Color and size flipped from 5
8[a,b]	40	70	0.15	31.2	24	4	Color flipped from 5
9[a,b]	70	40	0.15	24	31.2	4	Size flipped from 5

Treatments 5 and 6 are 2-dimensional. Treatments 7–9 are manipulations of treatment 5 designed to test the independence of perception of each dimension. Note that σ refers to the standard deviation (SD) for each dimension of phenotype. The size refers to the number of colored tiles (pixels) used to depict a prey item (linear dimension), with prey of (mean) sizes 24, 31.2, and 36 presented as sizes 1.3, 1.7, and 1.95 cm on the screen.
[a]"flipped" treatments
[b]only tested for relative cost = 1

exactly in the other dimension (see Table 1 treatments 1–4). For all four 1-dimensional treatments, we evaluated decisions under three different levels of cost.

Two-dimensional SDT

To address our first goal of evaluating how closely individual subjects' behavior could be modeled by SDT, we measured subjects' responses to model and mimic phenotypes over a range of costs and benefits when two dimensions could (and, in theory, should) be used to make the discrimination. First, we selected prey phenotypes from populations of models and mimics so that the relative information contained in size and coloration dimensions would be approximately equal (treatment 5). We also wanted to see how subjects behaved when there was more information in one dimension than another, so we conducted another treatment that was identical to treatment 5, except that size was more informative than color (treatment 6). Thus, in treatment 6, there was less difference in the mean color of model and mimic populations than in treatment 5, and more difference between their mean sizes, with variances held the same.

"Flipped" treatments

We needed to rule out the possibility that a correlation in the perception of color and size affected subjects' behavior when prey varied in two dimensions, for example, yellower prey might appear larger to subjects although, objectively, color and size vary independently (Ashby and Townsend 1986). To do this, we conducted a series of treatments where the mean colors and/or sizes of models and mimics were reversed with respect to treatment 5. Thus, if yellower prey did indeed appear larger to subjects, then we should be able to detect a difference in their behavior based on the thresholds adopted in the "flipped" treatments. There were three varieties of flipped treatments: one where color and size were reversed (treatment 7), one where only color was reversed (treatment 8), and one where only size was reversed (treatment 9).

Analysis

All models were fitted in R 3.1.3 (R Core Development Team 2015). First, we tested our assumption that presenting subjects with the distributions of models and mimics indeed educated them completely, that is, that they did not learn during trials. We did this by fitting a generalized linear mixed model to all of our data with the number of errors from a given number of attacks as the binomial response variable. We included group (each of the nine treatments from Table 1 divided into different levels of cost:benefit ratio) and the number of prey encountered as predictors, with individual participants allowed to have random slopes and intercepts. To test for a significant difference in error rate between groups, we compared this model to one that did not include group by using a likelihood ratio test. Based on our results (see Supplementary Figure 2), we discarded data from the first 20 prey items encountered for all subsequent analyses, as it appeared that some learning did occur early on as subjects encountered prey items and received feedback.

One-dimensional SDT

To evaluate whether size and color can each be used separately to categorize prey, we tested whether subjects' probability of attack was related to color or size when only one varied (i.e., in our 1-dimensional treatments 1–4), whether the difference between the mean phenotypes of models and mimics affected attack thresholds in a manner predicted by SDT, and whether or not costs of errors influenced subjects' behavior. Therefore, we fitted two generalized linear mixed models of attacks and rejections with the following predictors: color or size (continuous), the distance between the means of the model and mimic distributions (continuous), the relative cost of attacking models, and their interactions. The first model included subjects from treatments 1 and 2 (color informative), and the second included subjects from treatments 3 and 4 (size informative). In both models, subjects were allowed to have their own slopes and intercepts as random effects. Nonsignificant interaction terms were removed in a stepwise procedure.

Two-dimensional SDT

We wanted to know if subjects behaved in a manner consistent with SDT while discriminating in two dimensions, so we modeled each subject's probability of attacking a prey item as a function of the prey item's color and size. We considered five potential models for each subject: one in which subjects behave randomly, one with only prey size, one with only color, one with color and size, and another with color, size, and their interaction. All were generalized linear models with a logit link function, that is, binary logistic regressions. On fitting this type of model the subjects' decision boundaries can then be estimated (and plotted) as the 50% chance of attacking isocline (Figure 2). When subjects' behavior is best fit by a 1-parameter model that uses only size or only color, then the decision boundary is a horizontal or vertical line depending on what dimension is used for discrimination (Figure 2). When subjects use both color and size to a significant degree but not their interaction, their behavior is fit by a straight line that is neither horizontal nor vertical (Figure 2). The model that includes an interaction between color and size would have a curved decision boundary. Under SDT, a 2-parameter linear model describes the predicted optimal decision boundary under our experimental conditions (2-independent phenotypic dimensions). So, if individuals use both color and size to make decisions, then the model that best describes their behavior should be one that includes color and size but not their interaction. Note, however, that it is entirely possible for subjects to use linear decision boundaries that differ from the optimal slopes and/or optimal intercept. We used Akaike's Information Criterion (AIC) to select the best model for each human subject.

Often, individuals used only size or color as a predictor (see Results). We wanted to know if their use of either color or size depended on which experimental treatment was applied (i.e., color was approximately as informative as size, or size was more informative; treatment 6 vs. treatments 5, 7, 8, 9). To test the hypothesis that individuals randomly chose a single dimension to make decisions against the alternative that they preferred to use color unless size was much more informative, we used a Pearson chi-squared test. We applied this test to a table of treatment (6 vs. 5, 7, 8, 9) against subjects' use of color or size. Because expected counts of some cells were less than 5, the test statistic calculated by simulation using 99 999 replicates. We also used a generalized linear model with Poisson distribution (log link) to test whether the number of parameters in an individual's optimal model is related to treatment (i.e., how informative size was relative to color).

Pooling the behavior of individual subjects, we were curious to know how the behavior of a group of subjects affected attack rates on mimics, and hence selection for mimicry. For each group, we modeled the probability of attacking a prey item as a function of random decision-making (intercept only), prey color, prey size, color and size, or a maximal model that included color by size interaction, again using a generalized linear model with logit link function. Again, we used AIC to select the best model for each group.

We calculated the optimal decision boundary for each group under the 2-parameter model and tested whether the null hypothesis that the slopes and intercepts of group human decision boundaries in each treatment were different from the optimal ones. Although not all groups were best fit by the 2-parameter model, it was usually a candidate model (within 2 ΔAIC units of the best model; see Results), and it is convenient because it gives a decision boundary for the group that is linear in form, which can easily be compared with the analytical predictions of SDT. We asked whether the optimal slopes and intercepts fell within 95% of the bootstrapped slopes and intercepts. Rejection of either null hypothesis would indicate that attacks on mimics differed significantly from the predictions of SDT.

We examined the effects of costs of sampling and the relative information contained in each signal dimension. For this comparison, we analyzed only data from treatments 5 and 6. We fitted a generalized linear mixed model of the probability of attack on mimics with relative cost, color, size, treatment, and their interactions as additional predictors. Individuals were allowed to have their own slopes and intercepts with respect to color and size.

"Flipped" treatments

Finally, we needed to confirm that the perception and use of color and size in decision-making did not depend on particular values of color and size. Such an effect should be detectable by a shift in the decision-boundary for treatments where color and/or size were "flipped" for models and mimics. To test for this possibility, we fitted generalized linear models of the probability of attacking mimics with color and size of each mimic as predictors. We did this for each type of treatment: "regular," both color and size flipped, color flipped, and size flipped (treatments 5, 7, 8, and 9). We compared 95% confidence intervals (CIs) for the magnitudes of the slopes for the decision boundaries, which we calculated by bootstrapping. Our bootstrapping procedure created 1000 pseudoreplicates for each treatment dataset, with structure to account for individual subjects included. If flipped treatments did not alter subjects' behavior, then there should be no significant difference in the magnitude of their slopes.

RESULTS

Subjects encountered a mean of 163 ± 58 prey during the 5 min trials. A major assumption of SDT is that signal receivers (i.e., our human predators) already know the distributions of signal and noise (i.e., mimics and models) and act accordingly. We examined our data to detect a decrease in error rate with the number of prey subjects had encountered (Supplementary Figure 2). We also fit a logistic regression of subjects' error rate in classifying prey as a function of the number of prey they had encountered. Experimental groups differed in error rate committed (likelihood ratio test; $\chi^2 = 94.234$, df = 22, $P < 0.001$). This is expected because treatments vary in the overlap between model and mimic distributions, which determines discriminability. Although there was a slope to the regression line significantly different from zero (meaning that some learning occurred as humans encountered prey; Wald z test; $z = -2.156$, $P = 0.031$), its effect was slight, and little learning appeared to occur after the first 20 prey items that subjects encountered.

One-dimensional SDT

In our 1-dimensional trials, we found that humans in general used the appropriate dimension (color or size) to categorize prey (Wald z tests: color: $z = 3.51$; $P < 0.001$; size: $z = -7.82$; $P < 0.001$). With color, there was a significant interaction of the difference between means and color of individual prey ($z = -9.14$; $P < 0.001$), and the difference between means alone ($z = 7.54$; $P < 0.001$), but no effect of relative cost (see Table 2 and Supplementary Figure 3) on the attack rate of a given prey phenotype. In the case of size, both the difference between the means of the distributions ($z = 3.59$; $P < 0.001$) and the relative cost of committing errors ($z = -2.086$; $p = 0.03$) were significant predictors of the probability of attack a given phenotype (see Table 2 and Supplementary Figure 4).

Two-dimensional SDT

In two dimensions, individual subjects' behavior showed variation from the group mean for each treatment. Subjects were typically more variable in their behavior when size was more informative than color, that is, in treatment 6 rather than treatment 5 (Figure 4).

The crux of our assessment of how well human predators followed the predictions of multidimensional SDT was determining which model of attack probability best described their behavior. We found that the attacks made by individual subjects were usually best explained by single parameter models that contained either prey color or size (Figure 5). These subjects generally preferred to use color instead of size, except when size was much more informative (i.e., in treatment 6; exact chi-squared test, $P = 0.002$). However, we could not reject the null hypothesis that a few subjects guessed randomly, using no decision rule, while it appears that others used both size and color, or sometimes an interaction between the two (which implied they hypothesized a more complicated pattern than actually existed). In Figure 5, we also display the number of parameters in all models within 2 or 5 ΔAIC units of the best model. Often there was not a clear-cut best model for an individual, as can be seen from the initial rapid increase in the number of candidate models as higher ΔAIC values are included. The number of parameters in the models that best fit individual subjects' behavior was not significantly related to which treatment they experienced (Wald z-test; $z = -1.43$; $P = 0.15$), so more complicated discrimination rules

Table 2

Model formulas, coefficients, and standard errors, test statistics (Wald z), and significance values from 1-dimensional signal detection experiments

	Estimate	Standard error	z-value	P-value
Model: P (attack on mimic) ~ prop. blue • DBM				
Intercept	−2.54	1.27	−2.00	0.05
Prop. blue	8.09	2.31	3.51	<0.001
DBM	45.18	5.99	7.54	<0.001
Prop. Blue • DBM	−102	11.20	−9.14	<0.001
Model: P (attack on mimic) ~ relative cost + size + DBM				
Intercept	3.42	2.34	1.46	0.14
Relative cost	−0.26	0.13	−2.06	0.04
Size	−0.36	0.05	−7.82	<0.001
DBM	5.22	1.46	3.59	<0.001

DBM refers to difference between mean proportion blue of mimics and models, whereas prop blue refers to the proportion blue of any given prey type encountered. Individual subjects were treated as random effects in these binomial linear mixed models.

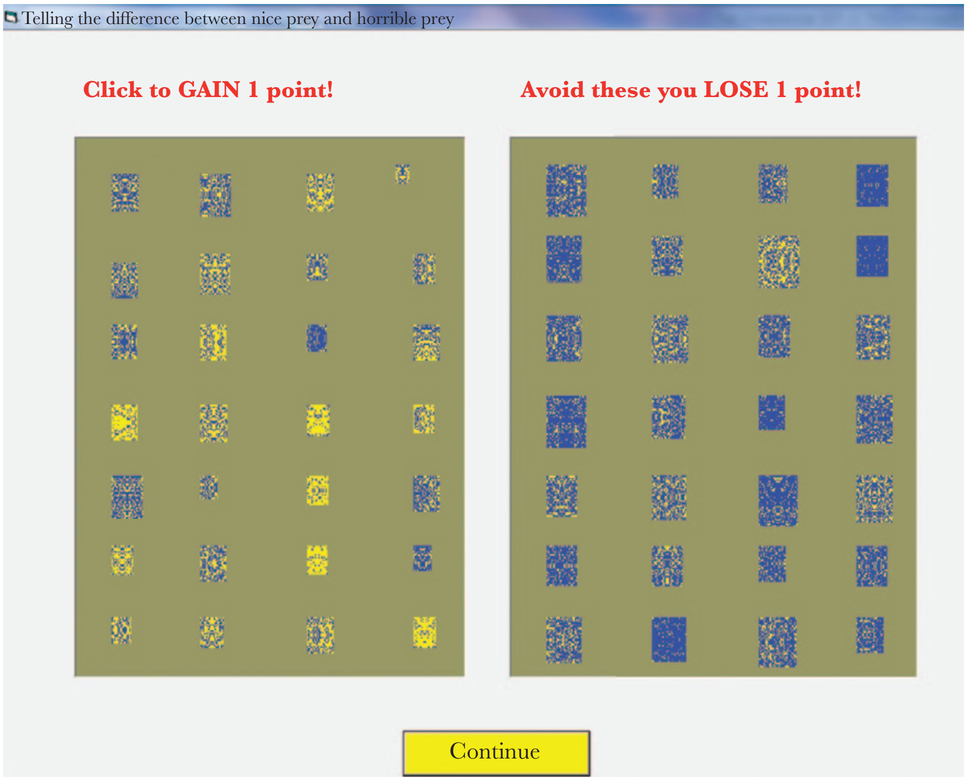
Telling the difference between nice prey and horrible prey
Click to GAIN 1 point!
Avoid these you LOSE 1 point!
Continue

Figure 3
Screenshot of the initial screen that subjects were shown prior to beginning. Here, mimics are on the left and models are on the right. Mean prop. blue of mimics = 0.4 ± 0.15, mean prop. blue of models = 0.7 ± 0.15. Mean size of mimics = 24 ± 4 pixels, mean size of models = 31.2 ± 4 (1.3 times bigger).

were not necessarily favored depending on whether or not size was more informative than color.

In contrast to individual subjects, attacks on a population of mimics by groups were usually best explained by more complex models (Figure 5). The best models most often included interactions, but it was often difficult to separate many of these from simpler, 2-parameter additive models (Figure 5d).

Although the behavior of a group of humans was sometimes best explained by a 2-parameter model that can contain the optimal decision boundary that is neither horizontal nor vertical, it does not in itself demonstrate optimal behavior because they could use both dimensions inappropriately. To better understand the relationship between attacks on mimics and the behavior of groups of subjects, we used bootstrapping to calculate CIs around slopes and intercepts of their decision boundaries. We found that only groups in treatments where size was much more informative than color had decision boundaries that contained either the slope or intercept of the appropriate optimal boundary, and even then only under some cost/benefit ratios (Table 3). The behavior of groups is most often explained by more complex models than the behavior of individuals (whose behavior is often best described by 1-parameter models—some of which use color and some of which use size), yet groups still usually deviate from the predictions of SDT.

Our model for predicting overall attack rates in two dimensions included size, color, treatment, and an interaction between color and treatment (Table 4). Most of the change in the decision boundary between treatments appeared to be mediated by a shift in the importance of prey color. In concordance with our other methods of analysis, when size becomes much more informative than color, color ceases to be used for decision-making.

"Flipped" treatments

Our bootstrap analysis showed that the slopes of decision boundaries between our "flipped" treatments 7–9 and the regular treatment 5 did not differ significantly (all simulated *P*-values > 0.05), indicating that dimensions were perceived independently.

DISCUSSION

We have shown that some qualitative predictions of SDT for mimicry may be borne out when applied to multiple dimensions, although the predictions are often not upheld when evaluated quantitatively. For example, as predicted, mimics are less likely to be attacked as they approach their models in phenotype, but the best-fit model for a group's behavior usually does not contain the optimal decision boundary—if it is even a straight line at all. This is broadly consistent with a similar experiment by McGuire et al. (2006), who found qualitative agreement with SDT but some quantitative deviations from optimality.

Individual subjects' behavior is not generally well-described by SDT. Many subjects focused on color while ignoring size in treatments 5, 7, 8, and 9, and in treatment 6 many appeared to use size exclusively although there was some information to be had in color (Figure 4). Even in 1-dimensional treatments, it was not uncommon for subjects' behavior to be best described by a random model rather than one of the form predicted by SDT. Although group behavior can be at least qualitatively described by SDT, this is an emergent property of the aggregate behavior of individuals, and could often be better represented by 3-parameter interaction models.

Why is human behavior not well explained by multidimensional SDT? Psychologists have extensively studied the performance

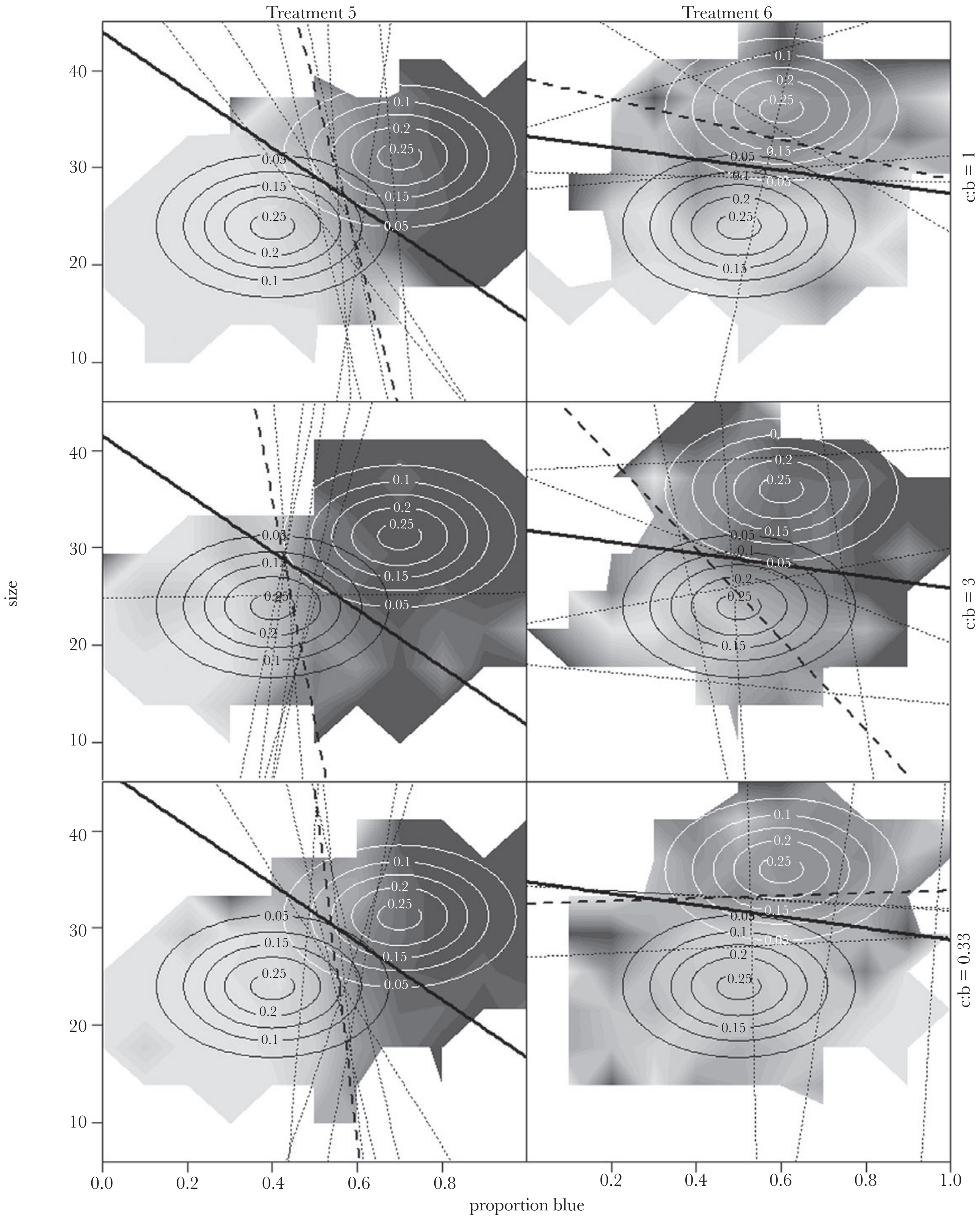

Figure 4

Reponses from groups of subjects in treatments 5 and 6. Densities of mimics and models presented are given by black and white contours, respectively. The optimal decision boundaries are shown by solid black lines, the group decision boundaries are shown by the dashed lines, and individual subjects' decision boundaries are shown by the dotted lines. All thresholds determined through the fit of a binary logistic model on attacking/not attacking prey with prey color and prey size as predictors, with the probability of attack set to 50%. The observed probability of attack decreases with darker shading.

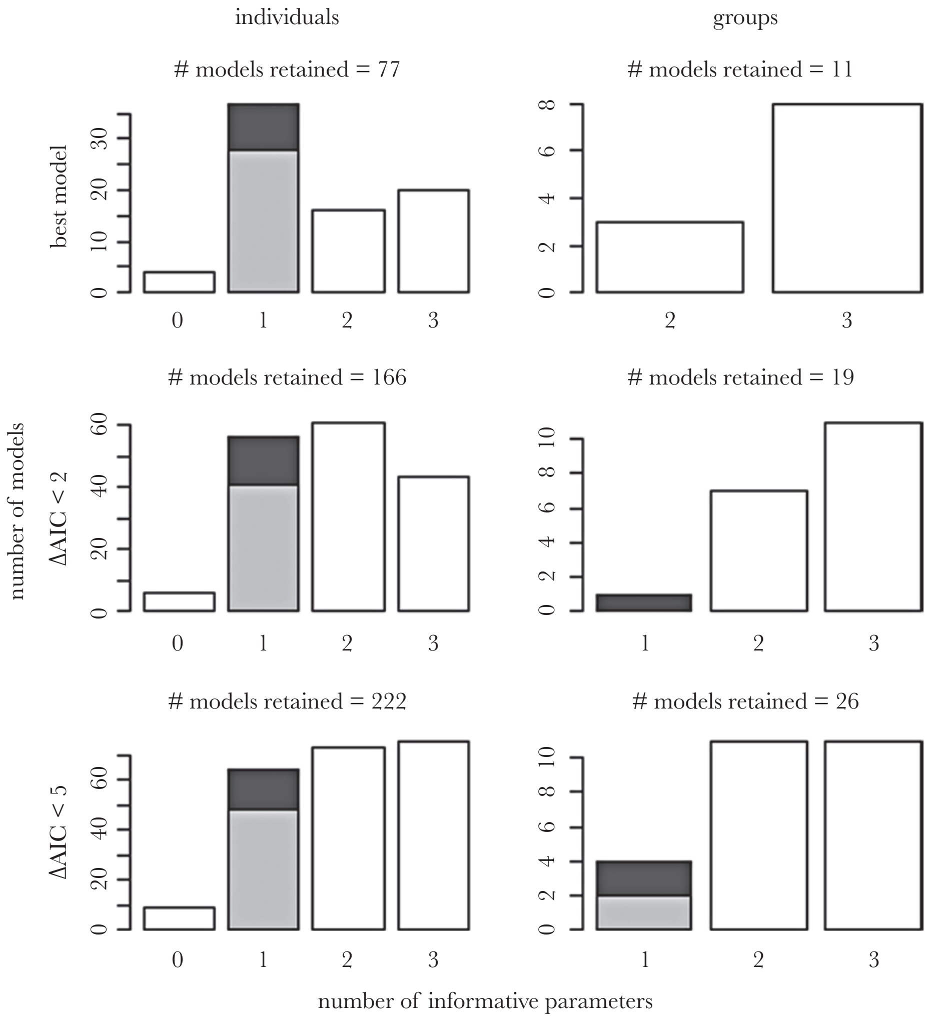

Figure 5

We fit models to the behavior of both individuals and groups for 2-dimensional SDT tasks (treatments 5–9) to find optimal decision boundaries. The models we considered were attack probability ~ 1, attack probability ~ color (one parameter; light gray), attack probability ~ size (one parameter; dark gray), attack probability ~ color + size (2 parameters), and attack probability ~ color * size (three parameters, including main effects). When 0-, 2-, or 3- parameter models provide the most parsimonious fit, neither or both phenotypic dimensions are employed (white). (a, c, and e) Distribution of the number of parameters for models fit to each of 77 individuals. (b, d, and f) Distribution of the number of parameters for the models fit to each of 11 groups. The top row (a and b) shows only the best models (lowest AIC); the middle row (c and d) shows all models within 2 ΔAIC of the lowest AIC, and bottom row shows all models within 5 ΔAIC.

of humans on multidimensional discrimination tasks for which the optimal solution is given by SDT. This body of modeling is often referred to as general recognition theory (GRT; Ashby and Townsend 1986; Ashby and Maddox 1990). Psychologists are interested in fundamentally different questions from behavioral ecologists; namely, they aim to describe the mental representations and decision-making processes of humans rather than explain behavior from the perspective of evolutionary optimality. Nonetheless, GRT is informative in predicting whether or not humans will follow the expectations of SDT, and has some implications for mimicry.

The most pertinent GRT concept for this experiment is that of decisional separability: does dimension A affect decisions made about dimension B across different levels of A? In our case, does the mean value of color affect how size is used to make decisions (i.e., the difference across treatments 5 and 6, but not variation within either treatment)? Clearly, it does, because when color and size are equally informative, subjects often use only color to make decisions. This happens despite our finding from 1-dimensional experiments that either color or size alone can be used to make discriminations. Therefore, we conclude that size is not entirely decisionally separable from color. Ashby and Maddox (1990) found a similar result when they asked subjects to discriminate between semicircles based on their size and the orientation of a line within them. When only one dimension contained information, subjects were capable of discriminating accurately, but when both were equally informative and independent, subjects preferred to use line orientation rather than size to classify stimuli.

There are many reasons why, in nature, size ought not to be a particularly valid cue in judging the properties of objects. First, any ambiguity in the distance of an object from the viewer will result in an inaccurate estimate of size. Size may also be uninformative for telling harmful prey from harmless prey; different species would frequently be expected to overlap in size during different stages of ontogeny. Color, on the other hand, will be much more informative due to mechanisms that enforce color constancy in the retina (Kelber et al. 2003). During learning experiments, animals in cognitive psychology experiments quickly learn to prefer one cue over another if it is a more valid predictor of a stimulus ("relative validity"; Shettleworth 2010). It seems likely that over evolutionary time, innate estimates

Table 3

Slopes and intercepts of decision boundaries for 2-dimensional trials on the group level, calculated from 1000 bootstrap pseudoreplicates

Treatment	C:B	Optimal slope	Empirical slope 95 % CI	*P*-value of slope	Optimal intercept	Empirical int. 95 % CI	*P*-value of intercept
5	0.33	1.60	1.57–1.58	<0.001	46.3	126 to 951	0.018
	1	1.60	1.57–1.58	<0.001	43.9	90 to 209	<0.001
	3	1.60	1.57–1.58	<0.001	41.5	82.2 to 343	<0.001
6	0.33	−1.40	−1.47 to 1.48	0.148	34.7	26.9 to 38.7	0.486
	1	−1.40	−1.52 to −1.18	0.212	33.3	33.9 to 45	0.028
	3	−1.40	−1.56 to −1.54	<0.001	31.8	40.1 to 60.9	<0.001
7	0.33	1.60	1.57 to 1.58	<0.001	41.5	104 to 249	<0.001
	1	1.60	1.58 to 1.59	<0.001	43.9	65 to 94.4	<0.001
	3	1.60	1.57 to 1.58	<0.001	46.3	138 to 1540	0.03
8	1	1.53	1.56 to 1.57	<0.001	11.3	−79 to −11.8	0.008
9	1	1.53	1.56 to 1.57	< 0.001	11.3	−78.1 to −11.5	<0.001

Two-tailed *P*-values reflect the probability that the 95% CI contains the optimal value. Treatments 7–9 are from "flipped" data where color and/or size were reversed from treatment 5, where color and size are approximately equally informative. Slopes are given in radians because sometimes the CIs include the *y* axis (1.571 or −1.571 radians).

Table 4

Table of fixed effects in our generalized linear mixed model of the probability of attack on mimics (binomial errors) with relative cost, color, size, treatment, and their interactions as additional predictors. Individuals were allowed to have their own slopes and intercepts with respect to color and size

	Estimate	Standard error	*z*-value	*P*-value
Intercept	23.30	5.50	4.24	<0.001
Relative cost	−4.45	2.63	−1.69	0.09
Prop. blue	−37.24	10.26	−3.63	<0.001
Size	−0.47	0.21	−2.21	0.03
Treatment	−17.76	6.72	−2.4	0.008
Relative cost • prop. Blue	5.70	5.31	1.07	0.28
Relative cost • size	0.16	0.10	1.49	0.13
Prop. blue • size	0.71	0.39	1.82	0.07
Relative cost • treatment	2.55	3.50	0.73	0.47
Prop. blue • treatment	35.72	12.32	2.90	0.004
Size • treatment	0.32	0.26	1.22	0.22
Relative cost • prop. blue • size	−0.25	0.21	−1.22	0.22
Relative cost • prop. blue • treatment	−3.15	6.74	−0.47	0.64
Relative cost • size • treatment	−0.08	0.14	−0.58	0.56
Prop. blue • size • treatment	−0.63	0.47	−1.34	0.18
Relative cost • prop. blue • Size • treatment	0.1	0.26	0.39	0.70

of relative validity would evolve if some cues are more consistently valid than others, for example, color versus size (Shettleworth 2005). Therefore, it should not be surprising that receiver psychology favors subjects' use of color over size when categorizing stimuli as costly or beneficial (Rowe 1999; ten Cate and Rowe 2007).

Relative validity is related to another concept from cognitive psychology called "overshadowing," which has also been invoked to explain why some traits on imperfect mimics are preferred over others (Cuthill 2014; Kazemi et al. 2014). In overshadowing, subjects presented with two cues that predict a stimulus form weaker associations between each cue and the stimulus than if they had been trained using single cues (Mackintosh, 1976). The effects of overshadowing are often asymmetrical, with one cue being more strongly associated with the stimulus than the other (Mackintosh 1976). Although overshadowing occurs during learning, whereas SDT assumes that subjects have complete knowledge, results from signal detection experiments such as those of Ashby and Maddox (1990) and this one suggest that subjects may consistently favor one dimension over another. If subjects have strong prior beliefs about the relative validity of cues such as color versus shape, size, or pattern, it could readily create overshadowing-like effects where subjects prefer to use one cue over another for categorization. Such effects are expected to have important implications for the origination of mimicry because its evolution could be initiated with a mutation that affects a single critical aspect of phenotype (Chittka and Osorio 2007; Gamberale-Stille et al. 2012), rather than a mutation of improbably large effect on all perceptible dimensions of phenotype (Punnett 1915; Nicholson 1927). Indeed, avian predators often use color instead of size, shape, or pattern when performing discrimination tasks (Terhune 1977; Kazemi et al. 2014), suggesting that color is a likely candidate trait to initiate the evolution of mimicry.

Generally speaking, we found that the effects of varying costs on subjects' behavior were quite weak relative to others. In a similar experiment on 1-dimensional computer prey that varied in color, McGuire et al. (2006) found that the proportion of mimics attacked was not significantly influenced by the relative abundance of mimics, but that the probability of an individual mimic being attacked was (these contrasting results were given by different methods of analysis). In our 2-dimensional treatment, we did not find that costs of errors significantly predicted attack probability. There are at least two potential explanations for why subjects' behavior is less sensitive to cost than expected. First, subjects might not take points in a computer game very seriously, and be more interested in figuring out the categorization task rather than optimizing their scores. Second, they might not assume that the phenotypes of prey are normally distributed and therefore form inaccurate estimations of where their decision boundaries should lie.

In sum, we have performed a test of 2-dimensional SDT as it applies to Batesian mimicry. Individual human subjects display wide variability in their decision boundaries, perhaps due to strong yet varying beliefs about the relative validity of size versus color for performing categorization tasks. Groups of subjects often also deviated from optimality. Studies of mimicry that involve continuous variation between models and mimics in multiple dimensions should consider the possibility that dimensions might not be weighted equally by predators. Color is of fundamental importance to discrimination ability, and could initiate the evolution of mimicry, but to better support such a conjecture, we need more studies of predators in natural mimicry systems and comparative information on the evolution of mimetic color patterns.

SUPPLEMENTARY MATERIAL

Supplementary material can be found at http://www.beheco.oxfordjournals.org/

FUNDING

This work was supported by the National Science Foundation (grant number NSF-1306037 to D.W.K.) and a Natural Sciences and Engineering Research Council Discovery Grant (T.N.S).

We would like to thank members of the Sherratt lab for their comments on this manuscript. This experiment was approved by the Carleton University Research Ethics Committee and was conducted according to the guidelines set out in Canada's Tri-Council policy statement on ethical conduct for research involving humans.

Handling editor: Johanna Mappes

REFERENCES

Ashby FG, Maddox WT. 1990. Integrating information from separable psychological dimensions. J Exp Psychol Hum Percept Perform. 16:598–612.

Ashby FG, Soto FA. 2015. Multidimensional signal detection theory. In Busemeyer JR, Townsend JT, Wang Z, Eidels A, editors. Oxford handbook of computational and mathematical psychology. New York: Oxford University Press. p. 13–34.

Ashby FG, Townsend JT. 1986. Varieties of perceptual independence. Psychol Rev. 93:154–179.

Bates HW. 1862. Contributions to an insect fauna of the Amazon valley (Lepidoptera: Heliconidae). Trans Linn Soc Lond. 23:495–556.

ten Cate C, Rowe C. 2007. Biases in signal evolution: learning makes a difference. Trends Ecol Evol. 22:380–387.

Chittka L, Osorio D. 2007. Cognitive dimensions of predator responses to imperfect mimicry? PLoS Biol. 5:2754–2758.

Cuthill IC. 2014. Evolution: the mystery of imperfect mimicry. Curr Biol. 24:R364–R366.

Dittrich W, Gilbert F, Green P, McGregor P, Grewcock D. 1993. Imperfect mimicry: a pigeon's perspective. Proc R Soc Lond Ser B-Biol Sci. 251:195.

Dorsi CJ, Swets JA. 1995. Variability in the interpretation of mammograms. N Engl J Med. 332:1172–1172.

Duncan CJ, Sheppard PM. 1963. Continuous and quantile theories of sensory discrimination. Proc R Soc Lond Ser. B-Biol Sci. 158:343–363.

Duncan CJ, Sheppard PM. 1965. Sensory descrimination and its role in the evolution of Batesian mimicry. Behaviour. 24:269–282.

Gamberale-Stille G, Balogh AC, Tullberg BS, Leimar O. 2012. Feature saltation and the evolution of mimicry. Evolution. 66:807–817.

Getty T. 1985. Discriminability and the sigmoid functional-response: how optimal foragers could stabilize model-mimic complexes. Am Nat. 125:239–256.

Goodale MA, Sneddon I. 1977. Effect of distastefulness of model on predation of artificial Batesian mimics. Anim Behav. 25:660–665.

Harper GR, Pfennig DW. 2007. Mimicry on the edge: why do mimics vary in resemblance to their model in different parts of their geographical range? Proc R Soc B-Biol Sci. 274:1955–1961.

Hebets EA, Papaj DR. 2005. Complex signal function: developing a framework of testable hypotheses. Behav Ecol Sociobiol. 57:197–214.

Holen ØH, Johnstone RA. 2004. The evolution of mimicry under constraints. Am Nat. 164:598–613.

Iserbyt A, Bots J, Van Dongen S, Ting JJ, Van Gossum H, Sherratt TN. 2011. Frequency-dependent variation in mimetic fidelity in an intraspecific mimicry system. Proc Biol Sci. 278:3116–3122.

Johnstone RA. 2002. The evolution of inaccurate mimics. Nature. 418:524–526.

Kazemi B, Gamberale-Stille G, Tullberg BS, Leimar O. 2014. Stimulus salience as an explanation for imperfect mimicry. Curr Biol. 24:965–969.

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

Kikuchi DW, Pfennig DW. 2010a. Predator cognition permits imperfect coral snake mimicry. Am Nat. 176:830–834.

Kikuchi DW, Pfennig DW. 2010b. High-model abundance may permit the gradual evolution of Batesian mimicry: an experimental test. Proc R Soc B-Biol Sci. 277:1041–1048.

Kikuchi DW, Pfennig DW. 2013. Imperfect mimicry and the limits of natural selection. Q Rev Biol. 88:297–315.

Leonard AS, Dornhaus A, Papaj DR. 2011. Flowers help bees cope with uncertainty: signal detection and the function of floral complexity. J Exp Biol. 214:113–121.

Lindström L, Alatalo R V, Mappes J. 1997. Imperfect Batesian mimicry—the effects of the frequency and the distastefulness of the model. Proc R Soc Lond Ser B-Biol Sci. 264:149–153.

Lynn SK. 2005. Learning to avoid aposematic prey. Anim Behav. 70:1221–1226.

Lynn SK, Barrett LF. 2014. "Utilizing" signal detection theory. Psychol Sci. 25:1663–1673.

Mackintosh NJ. 1976. Overshadowing and stimulus intensity. Anim Learn Behav. 4:186–192.

McGuire L, Van Gossum H, Beirinckx K, Sherratt TN. 2006. An empirical test of signal detection theory as it applies to Batesian mimicry. Behav Processes. 73:299–307.

Nicholson AJ. 1927. A new theory of mimicry in insects. Aust Zool. 5:10–104.

Oaten A, Pearce CE, Smyth ME. 1975. Batesian mimicry and signal detection theory. Bull Math Biol. 37:367–387.

Penney HD, Hassall C, Skevington JH, Abbott KR, Sherratt TN. 2012. A comparative analysis of the evolution of imperfect mimicry. Nature. 483:461–464.

Punnett RC. 1915. Mimicry in butterflies. London (UK): Cambridge University Press.

R Core Development Team. 2015. R: a language and environment for statistical computing.

Rowe C. 1999. Receiver psychology and the evolution of multicomponent signals. Anim Behav. 58:921–931.

Ruxton GD, Sherratt TN, Speed MP. 2004. Avoiding attack. New York: Oxford University Press.

Sherratt TN. 2002. The evolution of imperfect mimicry. Behav Ecol. 13:821–826.

Shettleworth SJ. 2005. Taking the best for learning. Behav Processes. 69:147–149.

Shettleworth SJ. 2010. Cognition, evolution, and behavior. Oxford: Oxford University Press.

Swets J, Tanner W, Birdsall T. 1961. Decision processes in perception. Psychol Rev. 68:301–340.

Terhune EC. 1977. Components of a visual stimulus used by scrub jays to discriminate a Batesian model. Am Nat. 111:435.

Wiley RH. 1994. Errors, exaggeration, and deception in animal communication. In: Real L, editor. Behavioral mechanisms in ecology. Chicago: University of Chicago Press. p. 157–189.