

A Hierarchical Bayesian Model of Individual Differences in Memory for Emotional Expressions

David Landy (dlandy@indiana.edu)

Indiana University, Bloomington

Department of Psychological and Brain Sciences, 1101 East 10th Street

Bloomington, IN 47405 USA

L. Elizabeth Crawford (lcrawfor@richmond.edu)

Jonathan Corbin (jcorbin@richmond.edu)

University of Richmond

Department of Psychology, 28 Westhampton Way

Richmond, VA, 23173

Keywords: category adjustment models; emotion memory; emotion perception; face perception; individual differences; Bayesian modeling

Abstract

When participants view and then reproduce simple objects that vary along a continuous dimension such as length or shade, or when they view images of faces that vary in emotional expression, their estimates tend to be biased toward the average value of the presented objects, a phenomenon that has been modeled as the result of a Bayesian combination of prior category knowledge with an imprecise memory trace (Corbin, Crawford & Vavra, 2017; Huttenlocher, Hedges & Vevea, 2000). Whereas previous work described a general cognitive strategy based on data aggregated across participants, here we examined individual differences in strategy. Thirty-six participants viewed and reproduced 496 morphed face stimuli that ranged from angry to happy. We found substantial variation in the bias patterns participants produced. Individuals' estimates were well fit by a model that posited attraction toward three categories, one at the happy end of the range, one at the angry end, and one that captured the entire range of presented stimuli, and by allowing the weight given to each category to vary by participant.

Introduction

Memories are never pure. Memory of an object is determined not only by that individual object, but also by the set, or category, to which it belongs. Specifically, items tend to be remembered as being more like the typical (average) item in a set than they actually were. For example, Huttenlocher, Hedges and Vevea (2000) had participants view and immediately reproduce individual items that varied along a continuous dimension such as length, width, or shade. They manipulated the presented distribution of lengths, widths, and shades and found that estimates were biased toward the central value of the distribution shown. They proposed the bias is a byproduct of a Bayesian combination of a noisy, unbiased memory trace of the stimulus with a prior distribution that reflects the presented stimuli. Related Bayesian accounts have been developed to

account for bias in time perception (Jazayeri & Shadlen, 2010), hue judgments (Olkkonen, McCarthy, & Allred, 2014), and estimates of the sizes of familiar fruits and vegetables (Hemmer & Steyvers, 2009). Here we extend this earlier work in two important ways. First, we apply this explanation to rich, socially relevant stimuli: faces that vary in emotional expression. Second, we model individual differences in how people rely on category knowledge when remembering facial expressions.

It is an open question whether memory for facial expressions can be characterized by the same principles that have been used to explain memory for length of a line. Facial expressions are socially meaningful and visually complex stimuli with which people have extensive prior experience, and unlike many other objects, faces are processed holistically (e.g., Maurer, Le Grand, & Mondloch, 2002). Compared to simple geometric objects, it is more difficult to assess visual memory of real faces. One approach is to use morphing software to create gradations of faces that vary along a dimension of interest. By morphing pictures of the same actor making angry, neutral, and happy faces, we can create a continuum of emotional expression to be used in memory tasks like the immediate reproduction procedure described above. These morphed continua allow researchers to assess the degree to which a particular face is remembered as having an expression that is more or less happy or angry than it actually was.

Few studies have used face morphs to examine bias in memory for individual facial expressions (but see Haberman, Brady & Alvarez, 2015; Haberman & Whitney, 2009 for related work). In a study designed to examine the central tendency bias in face memory, Corbin, Crawford and Vavra (2017) ran several experiments in which participants viewed faces one at a time and, after each one, estimated its expression by adjusting a response morph. Estimates were consistently biased toward the central value of the stimulus distribution, whether it ranged from very sad to neutral, very happy to neutral, or moderately happy to moderately sad. Furthermore, the degree of this central tendency bias

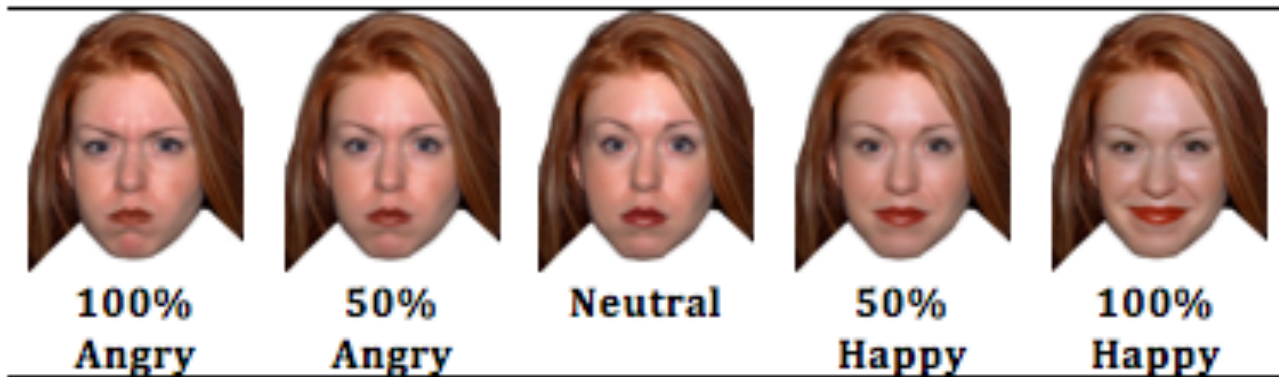


Figure 1: Example stimuli. Shown are the original angry neutral, and happy faces used to generate the stimulus morphs as well as morphed images between angry and neutral and between neutral and happy.

increased with longer retention intervals between stimulus and response. Bayesian models predict such an effect because, as the trace memory distribution becomes noisier (i.e., more variable), the Bayesian combination of trace memory and category knowledge will give more weight to the category knowledge (see also Huttenlocher et al., 2000; Crawford, Huttenlocher & Engebretson, 2000).

The Corbin et al. (2017) work was designed to allow for group-level conclusions and not for modeling of data from individual participants. This is typical of cognitive psychology, which usually characterizes the cognitive processing of a presumably generic, modal human mind without examining the variation between individuals. However, as we have noted elsewhere (Crawford, Landy & Presson, 2014; Crawford, Landy & Salthouse, 2016), that can lead to conclusions about aggregate tendencies that do not reflect the behavior or cognitive processing of any single individual. In fact, little is known about how people differ in their use of stimulus distributions to inform estimates of individuals. Building on the Corbin et al. findings, here we use Bayesian hierarchical modeling to examine both aggregate bias patterns and bias patterns at the level of individual participant. This approach allows us to estimate how each individual combines different category structures to arrive at estimates.

Emotional faces vary in physical dimensions such as mouth shape and brow orientation, as well as in affective significance, which can be processed automatically and unconsciously (e.g., Axelrod, Bar, & Rees, 2015; Vuilleumier, 2005). A continuum of emotional expression is necessarily bound up with physical feature variations and we do not attempt to tease these apart. Instead, we capitalize on previous work (Corbin et al., 2017; Haberman et al., 2015, Haberman et al., 2019) showing that the continuum created by morphing emotional faces produces results that mirror those found in studies using simple dimensions such as size, color, or shade. This work suggests that, when shown a set of faces that vary on a morphed expression continuum, people are sensitive to the central tendency of the set along that dimension.

Experiment

Method

Participants Thirty-six (11 male) students from the undergraduate participant pool at the University of Richmond received course credit for participating.

Materials Images were from the NimStim face stimulus set¹, a database of photographs of young adults depicting various emotional expressions. Sixteen models (8 male, 8 female) were chosen and the closed-mouth angry, neutral, and happy expressions of each were used to create the stimuli. Because in some cases, changes in hair position led to distracting artifacts in the morphed sets, we edited the initial images to maintain consistent hair placement. Using FantaMorph software (Abrosoft, 2002), each model's expressions were morphed from angriest to neutral to happiest, creating a set of 41 evenly distributed expressions that changed in 5% increments.

Procedure Each trial started with a crosshair at the center of the screen for 830 ms followed by a centrally presented single image frame taken from the morphed sets of faces and shown for 500 ms. The faces presented for study ranged from an angry expression (face #5) to happy (#35) and did not include the five most extreme images from either end of the continuum. After a blank screen (66 ms), a response face of the same model was shown in the upper left hand corner of the screen. Participants were instructed to "use the right and left arrow keys to change the expression of the face to match the expression of the previous photograph." Pressing the right arrow key made the expression cycle through the entire morph (images 0-40), cycling from happy to neutral to angry (or vice versa). Pressing the left arrow key cycled in the opposite direction. In a between subjects

¹ Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development. Please contact Nim Tottenham at tott0006@tc.umn.edu for more information concerning the stimulus set.

manipulation, participants were randomly assigned so that the starting frame of the response morph was always the angriest face (#0) or always the happiest face (#40). Participants estimated each of the 31 facial expressions for each of the 16 models, for a total of 496 randomly ordered trials.

Modeling

We modeled this data using a hierarchical Bayesian approach, simultaneously modeling individuals and group averages (see Figure 2). We assumed that each person was affected by a weighted combination of three potential biases: an overall inward bias toward the central category prototype (N), and two attractive biases toward postulated extreme categories, representing the endpoints of happiness (H) and anger (A). We assumed equal variance for each category, and a logistic categorization boundary. Each category had a separate ‘weight’ (W), which allowed the model to treat responses as the result of any number of categories from 0-3; best-fitting models uniformly predicted three categories (see Figure 5).

Explanations of bias are usually rooted in principles of Bayesian estimation: biasing responses toward a prior expectation reduces error (e.g., Feldman, Griffiths, & Morgan, 2009; Huttenlocher et al., 2000). In this initial analysis, we simply assumed that each category attracted responses toward its center. This structure captures the relationships most often studied in category-based adjustment experiments, but abstracts away from the relationship between variability and category use--components of the model which have previously met with some predictive success (Crawford et al, 2016), but which were to the side of our primary concerns in this initial analysis

Model predictions were unbounded, but actual responses were bounded between an extreme happy face (valued as 1), on one end of the scale, and an extreme angry face (-1). To handle this, we assumed that when participants retrieved a face beyond the edges of the scale, they would select the most extreme face available.

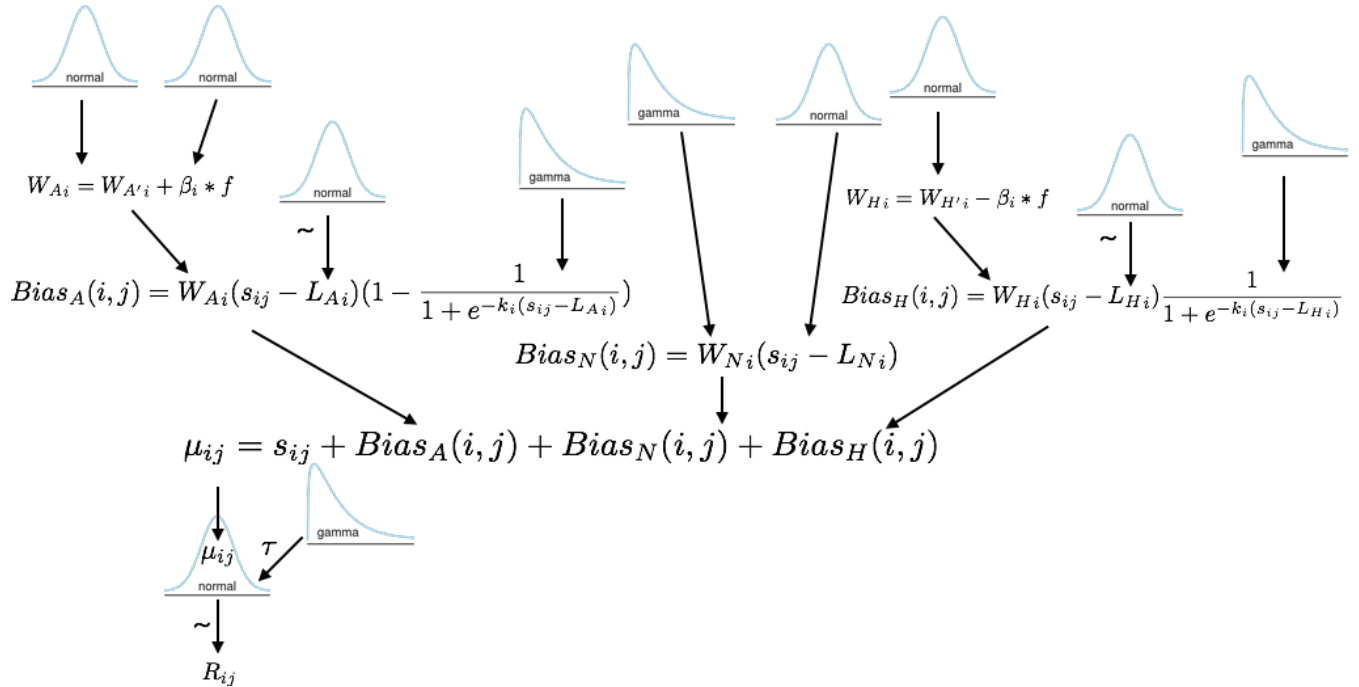


Figure 2: Graphical model diagram of the Bayesian model. R_{ij} is the response to stimulus j presented to subject i . The mean response is the sum of the stimulus value, s_{ij} , and three sources of bias, corresponding to the angry (A), neutral (N), and happy (H) prototypes. Each prototype has a weight (W) and a location (L). The category weights were potentially asymmetric, depending on the valence of the initial response slider (that is, whether the response face (f) was set to maximally angry (-1) or maximally happy (1). The model only shows the first layer of fits: all top-level distributions were governed by population-level hyper parameters (see Table 1), which employed weak priors. In all cases, we assumed unbounded parameters to be normally distributed, and positive unbounded parameters to be gamma.

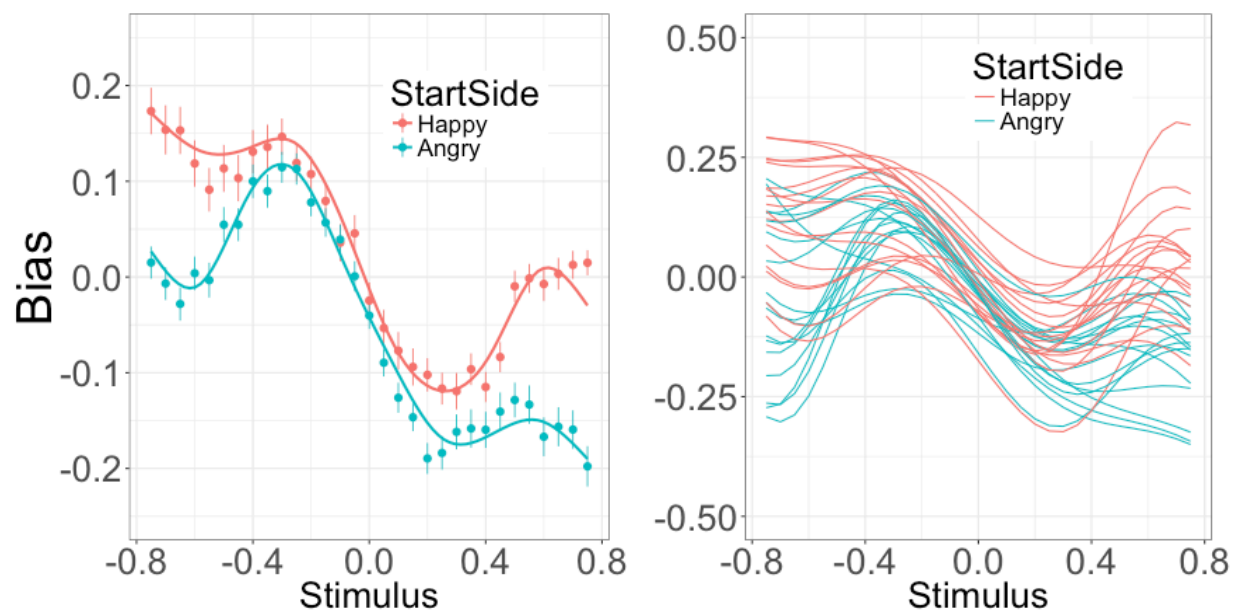


Figure 3: Aggregate and Individual Model Fits: (Left panel) Mean bias in response along with predictions averaged across participants. Errors reflect standard errors. (Right panel) Model fits for each individual participant. Use of all three categories is substantial, but starting side of the response strongly impacted the relative strength of these categories.

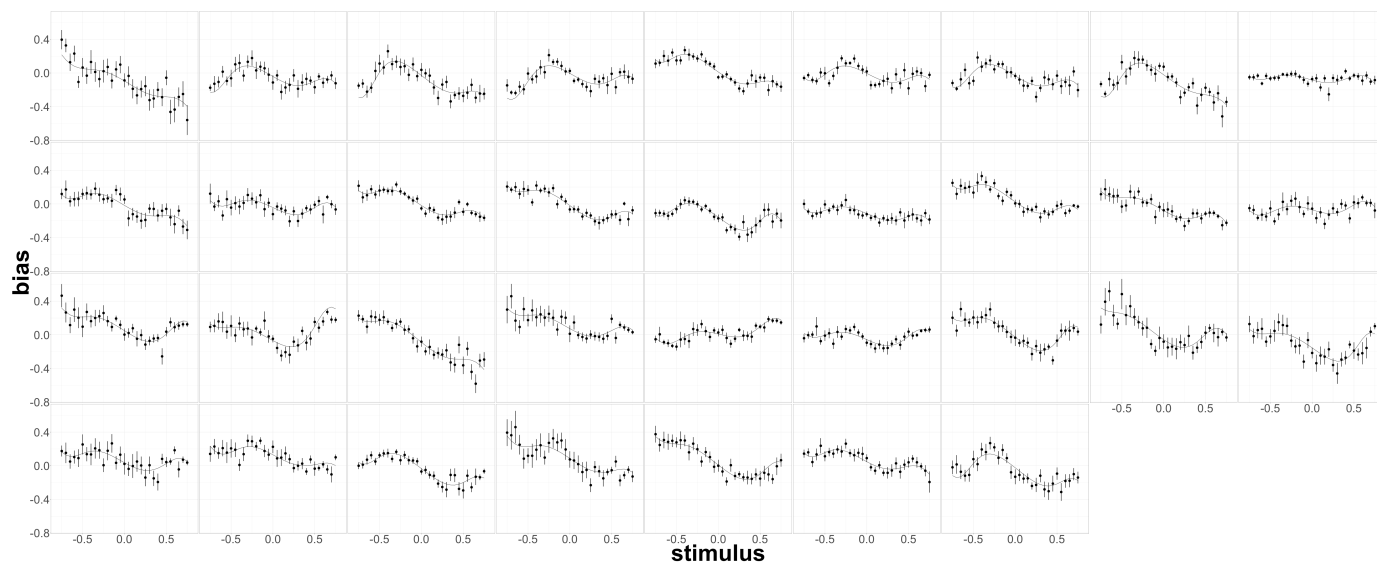


Figure 4: Individual model fits and data. Each dot is the bias in response to that stimulus, averaged across all times that participant viewed that expression. Each panel represents responses from one participant. Although different participants show quite different behaviors, the model treats each as a variation around a common theme of inward bias toward three weighted prototypes.

Results

Aggregated and individual response patterns are plotted in Figures 3 and 4. As can be seen, there was a strong pattern, overall, of attraction toward the center of the distribution. However, this was tempered by strong outward trends among most individuals. These outward biases tended to be moderately strong, roughly comparable in size to the bias toward the center, and in some cases dominating it. Figure 3 shows the aggregated model fits across participants; Figure 4 the individual fits.

Parameters fit hierarchically are listed in Table 1 and include the weights attributed to each category and the locations of each category.

The magnitude of the individual differences in weights can be characterized by the posterior deviation parameters (σ) governing weights. The 95% Highest Density Intervals for these excluded 0 (see Table 1), indicating that individuals differed in the weight given to these parameters (gamma shape parameters of roughly < 1 correspond to high density around 0), and that these differences were not well explained by sampling noise.

Table 1: *Priors and posteriors of population parameters. The μ values on the locations indicate mean locations of the categories, while the weight parameters have shape and rate values.*

Parameter	Population Prior	95% HDI
W_A	$\Gamma(\text{shape}, \text{rate})$ $\text{shape} \sim \Gamma(1, 0.005)$ $\text{rate} \sim \Gamma(1, 0.005)$	shape: [11, 168] rate: [6, 92] mean: [1.6, 2.4]
W_N	$\Gamma(\text{shape}, \text{rate})$ $\text{shape} \sim \Gamma(1, 0.005)$ $\text{rate} \sim \Gamma(1, 0.005)$	shape: [39, 196] rate: [42, 175] mean: [.85, 1.2]
W_H	$\Gamma(\text{shape}, \text{rate})$ $\text{shape} \sim \Gamma(1, 0.005)$ $\text{rate} \sim \Gamma(1, 0.005)$	shape: [120, 235] rate: [57, 124] mean: [1.6, 2.6]
L_A	$N(\mu, \tau)$ $\mu \sim N(-1, 80)$ $\tau \sim \Gamma(1, 200)$	$\mu: [-1.3, -1.15]$ $\tau: [.2, 2640]$
L_N	$N(\mu, \tau)$ $\mu \sim N(0, 80)$ $\tau \sim \Gamma(1, 200)$	$\mu: [-0.06, -0.002]$ $\tau: [68, 560]$
L_H	$N(\mu, \tau)$ $\mu \sim N(1, 80)$ $\tau \sim \Gamma(1, 200)$	$\mu: [1.225, 1.325]$ $\tau: [52, 2040]$
τ_{au}	$\Gamma(\text{shape}, \text{rate})$ $\text{shape} \sim \Gamma(3, 1)$ $\text{rate} \sim \Gamma(3, 1)$	$\text{shape: [3600, 10000]}$ rate: [240, 520]
β (side bias)	$N(\mu, \tau)$ $\mu \sim N(0, 80)$ $\tau \sim \Gamma(5, 0.1)$	$\mu: [0.01, 0.02]$ $\tau: [4,800, 28800]$

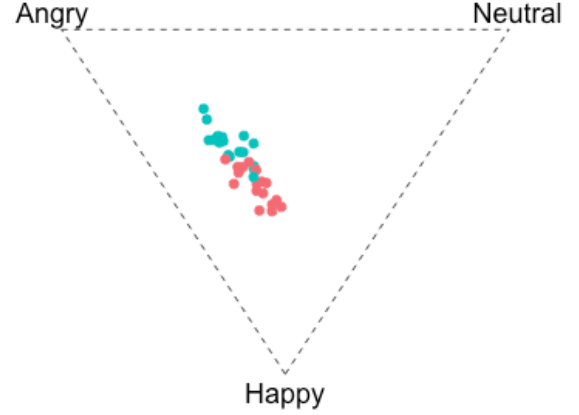


Figure 5: **Simplex plot of the relative weights accorded to each category.** A dot reflects a mean individual. Red indicates starting values on the happy side, blue on the angry side. Although in principle, the total weight could vary, in practice each individual showed a mean weight between 3.3 and 3.5, making simplex plots a useful visualization of the three values.

One factor had a strong apparent impact on the weight given to the left and right categories: the starting location of the response. To quantify this effect, we modeled the left and right weights as symmetric, except for a mean shift determined by an individual splitting parameter. This splitting parameter was fit to individuals; the posterior fits are shown in Figure 5. The results suggest a moderate impact of start location on category weight, such that people more heavily weighted the category represented in the starting value.

Discussion

Building on earlier work on inductive category effects on memory, we assume that estimates of an individual object combine an inexact memory trace of the object with knowledge of the set to which it belongs (e.g., Huttenlocher et al., 2000), producing estimates that are biased toward category prototypes. Such central tendency effects have been shown in studies using immediate reproduction tasks with simple stimuli that vary on one or two dimensions, such as size and shade (e.g., Crawford et al., 2001). Extending this work to more complex and socially relevant stimuli, Corbin et al. (2017) found that estimates of emotional expressions are also biased toward the center of the presented range of expressions, suggesting that participants used an inductively formed category to adjust estimate of faces.

Here we further examined the kinds of category structures involved in face memory and the degree to which individuals differed in their use of these structures.

As in previous work, estimates generally were biased toward the presented distribution's center (here a neutral expression). In addition, we found substantial variability between participants such that most participants were not well described by a model that treated estimates as resulting from adjustment toward a single, centrally located category. Good model fits at the participant level were achieved by positing that estimates could be adjusted toward two additional categories (centered on angry and happy values) and by allowing category weights to vary by participant. We note that this three-category model reflects the structure that was used to generate the stimuli: pictures of faces that actors made when told to show happy, angry and neutral expressions.

Some of the difference in how participants weighted the different categories could be accounted for by the starting value of the response face, which was randomly assigned between subjects. On average, greater weight was given to the category that aligned with the starting position (either 100% happy or 100% angry). The effect of the starting value was not linear across the stimulus range, as would be expected by inadequate adjustment away from an anchor. Instead it appears that the starting value encouraged participants to rely more heavily on the closest emotion category. Although studies of inductive category learning typically focus on the distribution of test objects, this result suggests that response objects may also contribute to the category structure used during estimation.

It is common to analyze group-level data and describe the collective's average behavior, but this approach can miss meaningful variation in cognitive strategies used by individuals. Modeling responses at the individual level reveals similarities across participants as well as some systematic differences. From the current study, it is not known why people adopt the strategies that they do. The model's success in capturing the different data patterns produced by individuals makes it a valuable framework for future studies of how differences in cognitive, social, and affective processing may influence the reliance on categories when remembering emotional faces. The variation in bias that we observed suggests that models pitched at the level of group averages are likely to mislead us away from the best interpretations.

References

Axelrod, V., Bar, M., & Rees, G. (2015). Exploring the unconscious using faces. *Trends in cognitive sciences*, 19(1), 35-45. doi: 10.1016/j.tics.2014.11.003

- Corbin, Crawford & Vavra (2017). Misremembering emotion: Inductive category effects for complex emotional stimuli. *Memory & Cognition*, pp. 1-8. doi: 10.3758/s13421-017-0690-7
- Crawford, L. E., Huttenlocher, J., & Engebretson, P. H. (2000). Category effects on estimates of stimuli: Perception or reconstruction? *Psychological Science*, 11, 280-284. doi:10.1111/1467-9280.00256
- Crawford, L. E., Landy, D., & Presson, A. N. (2014). Bias in spatial memory: Prototypes or relational categories? Proceedings of the 36th Annual Conference of the Cognitive Science Society. Quebec City, Quebec: Cognitive Science Society.
- Crawford, L. E., Landy, D., & Salthouse, T. A. (2016). Spatial working memory capacity predicts bias in estimates of location. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9), 1434-1447. doi: 10.1037/xlm0000228
- Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4), 752-782. doi: 10.1037/a0017196
- Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, 144(2), 432-446. doi: 10.1037/xge0000053
- Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 718-734. doi: 10.1037/a0013899
- Hemmer, P., & Steyvers, M. (2009). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review*, 16(1), 80-87. doi: 10.3758/PBR.16.1.80
- Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, 129, 220-241. doi: 10.1037/0096-3445.129.2.220
- Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, 13, 1020-1026. doi:10.1038/nn.2590
- Maurer D., Le Grand R., Mondloch C.J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, 6, 255-260. doi: 10.1016/S1364-6613(02)01903-4
- Olkkonen, M., McCarthy, P. F., & Allred, S. R. (2014). The central tendency bias in color perception: Effects of internal and external noise. *Journal of Vision*, 14, 1-15. doi: 10.1167/14.11.5
- Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences*, 9(12), 585-594. Doi: 10.1016/j.tics.2005.10.011