Two paths to variation in semantic judgments: How ambiguity and conceptual diversity drive individual differences in meaning

Di Liu[1] **(dliu88@jh.edu), Raja Marjieh**[2] **(raja.marjieh@princeton.edu),**
Nori Jacoby[3] **(kj338@cornell.edu), and Robert Hawkins**[4] **(rdhawkins@stanford.edu)**
[1]Department of Psychological & Brain Sciences, Johns Hopkins University,
[2] Department of Psychology, Princeton University,
[3]Department of Psychology, Cornell University,
[4]Department of Linguistics, Stanford University

Abstract

Why do individuals differ in the way they assign semantic labels to the same perceptual referent? One possible source of disagreement is referential ambiguity, where stimuli near category boundaries are harder to label. Another is latent diversity in conceptual representations, leading to labeling differences even for confidently categorized referents. To distinguish between these sources of variation, we used Gibbs Sampling to search through the multidimensional Chernoff face space to find faces that are prototypically happy or sad, or ambiguous (Experiment 1, $N = 253$). Then in Experiment 2 ($N = 684$), we asked a naive group of participants to rate the emotions of these faces, finding that ambiguous faces elicited greater individual differences in valence interpretation and a medium level of variation when being labeled using basic emotional terms. Simultaneously, even well-categorized happy and sad faces triggered variability in their consensus labels, though showing less disagreement when mapped onto an obviously inappropriate label. These findings suggest that both categorical boundaries and within-category variability shape individual differences in semantic interpretation.

Keywords: Individual Difference; Perceptual Representations; Gibbs Sampling; Emotion Perception

Introduction

When people in the same community talk about the things around them, there's a remarkable degree of consensus in the labels they use. We can readily agree that we're sitting in 'chairs', point to the same patch of sky when discussing 'blue', and identify the same 'happy' facial expressions. Yet beneath this apparent coordination lie subtle but profound differences in how we mentally represent what words mean. Like the topiary gardens evoked by Quine (2013) – where the visible outer branches are neatly trimmed to match but the inner growth remains wild and varied – our semantic representations achieve surface alignment through communication but may harbor hidden variations in their deeper structure. Recent work in cognitive science has increasingly recognized this latent diversity (Botch & Finn, 2024; Feldman & Choi, 2022; Johns, 2024; Marti et al., 2023; Oktar et al., 2024; Sarafoglou et al., 2024; Wang & Bi, 2021). For example, Wang and Bi (2021) found that individuals show great disagreement in representing word meanings, particularly in abstract words, such as "relationship" or "scenery". However, it remains unclear what lies at the root of these hidden differences, or how to reliably surface them in the lab.

One possible source of observed individual differences arises from *referential ambiguity* – when exemplars fall in a fuzzy, uncertain region between clear categories. A face that blends features of both happiness and sadness, for instance, may be categorized differently by different people if they were forced to pick just one label. Like a color that lies somewhere between blue and green, these boundary cases naturally elicit disagreement because they lack clear category membership. Critically, referential ambiguity is equally ambiguous for *everyone*, leading to low confidence and within-subject variability in labeling decisions. A second, more subtle source of individual differences stems from *conceptual diversity* – differences in the mental representation of the underlying concepts themselves. These representational differences can manifest in various ways, from variations in how strongly people judge a prototypical happy face as 'happy', to more dramatic cases where individuals fundamentally disagree about whether a clearly categorized stimulus belongs in the category at all.

Distinguishing referential ambiguity from conceptual diversity is crucial because these sources of individual differences may require different strategies for resolving miscommunications (e.g. Duan & Lupyan, 2023; P. Li et al., 2017; Murthy et al., 2022). To do so, it is crucial to examine a semantic domain that supports both types of variability. In this paper, we focus on the domain of basic emotion concepts, specifically whether faces are *happy* or *sad*. This domain is particularly well-suited for our investigation for three key reasons. First, the recognition of basic emotions has been argued to have high consistency across individuals (e.g. Ekman et al., 2013), although the degree of universality remains controversial (e.g. Nelson & Russell, 2013). Second, emotional experiences are inherently subjective, and prior research suggests that people's conceptual representations of emotions may still exhibit individual variations (Brooks & Freeman, 2018; Winter & Kuiper, 1997). Third, faces exist along a perceptual continuum that should include ambiguous emotions that do not clearly belong to either end point.

To systematically manipulate and assess variations in emotion labeling, we use *Chernoff faces*, a parametric model that allows continuous control over eight distinct facial features of a cartoon face, such as curvature of the mouth, eye shape, and brow positioning, providing a robust foundation for generating a spectrum of emotional expressions (Chernoff, 1973). This structured stimulus space enables us to identify faces that are perceived as *ambiguous* in a probabilistic emotion

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

representation (e.g., ambiguously happy), or as *highly probable* exemplars of emotion concepts (e.g., consensually happy or sad), and to explore how individuals differ when interpreting these faces.

Experiment 1

We begin by using an approach known as *Gibbs Sampling with People* (GSP; Harrison et al., 2020; Van Geert & Jacoby, 2024) to search over the large space of possible faces and discover targets that are rated highly on happiness or sadness, as well as targets explicitly rated as highly ambiguous between the two. In practice, participants iteratively adjusted one of the facial dimensions using a continuous slider to optimize for one of three criteria (*sad*, *happy*, and *ambiguous*), allowing us to efficiently surface stimuli that may elicit individual differences. This Gibbs Sampling technique critically does not simply search for a single global optimum, but samples from the population distribution, and is thus well-suited for identifying sources of variation.

Methods

Participants We recruited $N = 253$ subjects from Prolific (131 female). Each subject gave consent at the beginning of the study, and received compensation based on the total time they spent on the task. All subjects gave consent to participate in the study.

Stimuli We used the visual stimulus space of *Chernoff faces* – a parametric model allowing independent manipulation of eight facial features: eyebrow shape, eye width, eye height, nose width, nose height, facial width, hair shape, and mouth shape. This constitutes a continuously varying, multidimensional space of possible faces that vary systematically in their appearance, and flexibly express a spectrum of emotions. Each face was presented on the participants' desktop or laptop computer within a 300×300 pixel window, centered at the top of the screen. Below the face, participants used a slider to dynamically adjust a given facial dimension. The slider always began at the center of the scale and manipulated the facial feature of the displayed face in real-time. This interactive design enabled participants to efficiently explore and optimize the face according to their assigned criterion.

Procedure Each participant was randomly assigned to one of three conditions determining their goal: using the slider to optimize *happiness*, optimize *sadness*, or optimize *ambiguity*. In the happiness and sadness conditions, participants were asked to make the face appear as *happy* or *sad* as possible, respectively. In the ambiguity condition, participants were asked to maximize their uncertainty about whether the face was happy. To ensure they understood the concept of ambiguity, participants were provided with a brief definition before the trials began ("Some faces look really happy. Some look really sad. Others look angry or surprised... But for some faces you're not really sure. They're ambiguous.").

Unbeknownst to the participants, they were collectively participating in a Gibbs sampling chain (see Figure 1A). This method has proven effective for mapping how populations perceive the referents of semantic labels, such as pleasantness in Harrison et al., 2020. Each task condition was structured into multiple parallel sampling chains that evolved through the following steps:

1. **Initialization:** Each chain began with a randomly generated face, where all eight facial dimensions were assigned random values.
2. **Sequential Adjustment:** This face was shown to a participant, who adjusted one randomly selected facial feature according to their assigned goal (i.e. happy, sad, or ambiguous).
3. **Propagation:** The adjusted face was passed to the next participant in the chain, who modified a different feature. This process continued until all eight features had been adjusted once (completing one "iteration cycle"). Each chain went through 80 iterations total, comprising 10 full cycles.

We ran 16 independent chains for the happiness condition, 10 for the sadness condition (collected earlier), and 16 in the ambiguous condition, yielding 3360 total faces across all chains and conditions, reflecting both ambiguous faces and prototypical exemplars of happiness and sadness. Experiment and analyses code can be found at: https://github.com/AuroraLiuDi/TwoIndividualDiff.git

Results

We first validated whether our Gibbs Sampling chains successfully converged by examining how faces from different conditions were distributed in the multidimensional feature space. A successful optimization should show two key patterns. First, the faces from the happiness and sadness conditions should occupy distinct regions, reflecting their opposing valence. Second, the ambiguous faces should fall between these regions, representing stimuli that are less prototypically (but not impossibly) perceived as happy or sad.

To visualize these distributions, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) to project the 8-dimensional Chernoff face features onto a 2D space. Given the iterative nature of Gibbs Sampling, we focused on analyzing a subset of faces that captures the progression of each chain without involving too much redundant information. This selected subset included: (1) 1 face representing the initial starting point of each chain, (2) 8 faces from the first iteration round, (3) 4 faces from the second iteration round, and (4) 2 faces from each of the subsequent eight cycles of iterations. This selection process ensured that our final stimulus set captured both the early-stage and progressively refined representations of each condition. A total of 138 faces were dropped due to missing values in the stored data. Finally, we compiled a set of 1028 faces.

As shown in Figure 1B, faces optimized under different criteria formed distinct clusters, with ambiguous faces occupying an intermediate space. A Permutation Multivariate Analysis of Variance (PERMANOVA) on the position of faces in

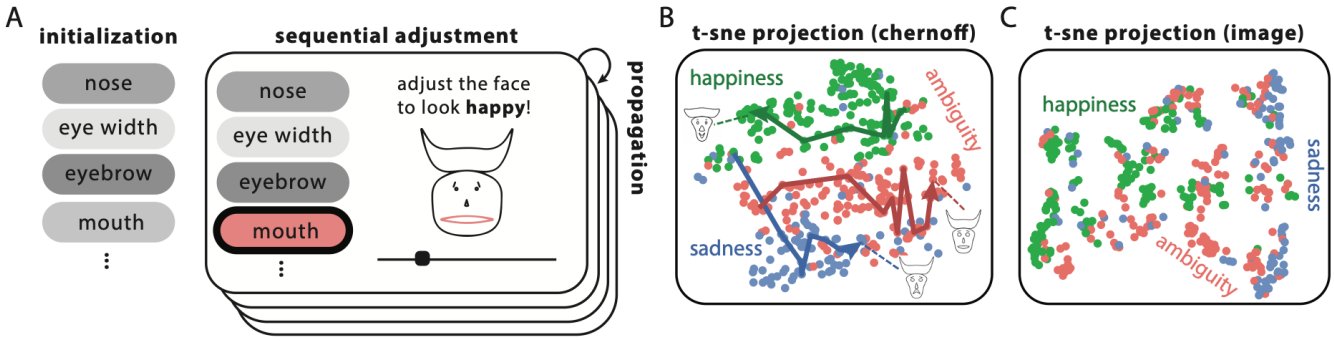

Figure 1: (A) The Gibbs Sampling procedure elicits distributions of faces across conditions (happiness condition shown). The faces from these chains are shown projected into two dimensions (B) from the Chernoff feature space and (C) from high-level image features. Examples of optimized faces are shown in (B). Arrows indicate the trajectory of example chains.

the original 8-dimensional space confirmed that faces generated under different conditions were reliably distinguishable ($F(2,1205) = 194.58$, $p = .001$). These results suggest that participants could consistently access and manipulate both prototypical examples and boundary cases associated with emotional labels.

To ensure these clusters were not an artifact of the 8-dimensional feature space of the Chernoff face parameterization, we conducted a secondary analysis based on high-level visual features directly extracted from rendered images of the faces. We used ResNet-18, a convolutional neural network trained on large-scale visual data, to obtain a $512 \times 7 \times 7$ feature matrix for each image using the second to last layer in the model – this layer was chosen to reflect the higher-order semantic features of the images as opposed to low-level visual features. When projected using t-SNE, we found the same basic pattern. As shown in Figure 1C, happy and sad faces remained well separated in the projected feature space. Interestingly, ambiguous faces clustered closer to happy faces in this space, but still formed a distinguishable group (pairwise comparison: $F(1,926) = 100.32$, $p = .001$; due to the large size of the raw feature matrices, we conducted the PERMANOVA on the positions of faces in the 2-dimensional t-SNE space for tractability). This suggests that ambiguous faces share key visual properties with prototypically happy faces, but remain distinct, exactly what we would expect from a successful optimization procedure: they appear to be less probable edge cases of the happy category, rather than entirely non-happy faces.

Experiment 2

Experiment 1 successfully generated three distinct sets of faces: those optimized to be perceived as happy, those optimized to be perceived as sad, and those optimized to be perceived as ambiguous. In Experiment 2, we had a naive group of participants rate these optimized faces on multiple emotion dimensions. These distinct distributions provide an ideal testbed for distinguishing our two hypothesized sources of individual differences. If variation in semantic judgments stems primarily from referential ambiguity, we should see greater individual differences when people rate faces that were explicitly optimized for ambiguity compared to consensus, and individual participants should show more uncertainty (within-subject variability) when rating ambiguous faces. If conceptual diversity plays a significant role, we should also expect to see systematic individual differences for consensus-optimized faces, and these judgments should be more confident.

Methods

Participants We recruited $N = 684$ subjects from Prolific. Each subject received compensation based on the total time they spent on the task. All subjects gave consent before participating in the study.

Stimuli The stimuli for Experiment 2 were the set of 1028 faces selected in the Experiment 1 analyses.

Procedure Each participant was presented with 20 randomly selected faces from the stimulus set and asked to evaluate the emotions expressed by each face. Specifically, they rated the extent to which each face conveyed Happy, Surprise, Sad, Angry, Fear, and Disgust using six separate sliders. Each slider ranged from "Least" (left end) to "Most" (right end). The initial positions of the sliders were hidden at the beginning of each trial to avoid priming effects. Participants' ratings were recorded as continuous values ranging from -1 to 1 for each emotion label. To ensure a robust evaluation of each face, a minimum of five participants rated each stimulus, with most faces receiving evaluations from ten participants. Here, the evaluation of emotion intensity reflects participants' judgments on how well the presented face maps onto given semantic labels.

Analytic approach

Preprocessing the data. We applied a preprocessing pipeline to ensure data quality, avoiding potential distortions.

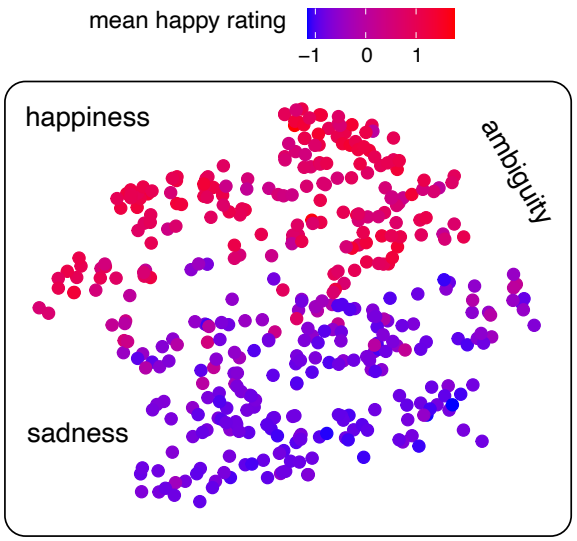

Figure 2: t-sne projection of faces from Figure 1B colored by mean happiness rating elicited in Experiment 2.

The following steps were applied: (1) if a participant rated a face as the least fitting ($< -.90$) or most fitting ($> .90$) across *all* six emotion labels, their ratings for that face were excluded from further analysis, (2) to control for participants' varying use of the slider, each participant's ratings on each basic emotion were z-transformed (standardized), (3) for each emotion dimension, ratings that deviated beyond 1.5 times the interquartile range (IQR) from the quartiles of all ratings were identified as extreme outliers. 1014 ratings (6.96% of all) were excluded for this reason. In addition, ratings on 6 faces were not included due to missing values.

Computing averaged ratings for each face. To evaluate the reliability of our ratings and the effectiveness of the Gibbs Sampling approach, we calculated the mean rating for each face across all participants for the labels of happy and sad. This enabled us to verify whether faces optimized in the happiness condition were, on average, rated as strongly representative of the "happy" label while being rated lower on "sad", and vice versa for the sadness condition. We also aimed to explore how the ambiguously happy faces are interpreted on average.

Selecting methods to calculate individual differences. A central goal of Experiment 2 was to assess how the faces mapped in Experiment 1 elicited individual differences in semantic interpretation. We employed two approaches to quantify the extent of variability across participants:

1. **Variation in valence perception:** Considering that ambiguously happy faces are prominently ambiguous referents in the domain of valence, we examined if individuals vary more in interpreting their valence compared to that for consensually categorized faces. We consolidated their ratings on six separate basic emotions into positive and negative scores. Specifically, their ratings on Happy and Surprise dimensions were averaged to generate a score on positive, while the averaged ratings on Sad, Angry, Fear, and Disgust reflected the score of negative valence. Inter-subject variability was quantified using Euclidean distance between pairs of participants' valence ratings for the same face (similar to that used in Z. Li et al., 2023). A higher mean distance across participants indicated greater disagreement in valence interpretation.

2. **Variability in specific emotion labeling** To examine individual differences in assigning specific basic emotion labels (with a focus on Happy and Sad), we computed both the standard deviation (SD) and the IQR of ratings for each label per face. Given the relatively small number of ratings per face, we applied a bootstrap resampling approach with 100 iterations. This ensured that variability estimates were not unduly influenced by potential outliers, providing more reliable measures of individual differences.

Results

Mean ratings across chain progression and conditions. We first examined how Happy and Sad ratings evolved over successive iterations within each chain (Figure 3A). As predicted, in chains optimizing happiness, faces were rated as significantly more Happy ($\beta = .53$, $SE = .05$, $t = 11.58$, $p < .001$) and less Sad ($\beta = -.38$, $SE = .04$, $t = -10.48$, $p < .001$) immediately after the first full cycle of iterations. Similarly, in chains optimizing sadness, faces were rated as significantly more Sad ($\beta = .57$, $SE = .09$, $t = 6.22$, $p < .001$), and less Happy ($\beta = -.24$, $SE = .07$, $t = -3.45$, $p < .001$). Converging evidence showed that faces from the happiness chain were rated as more Happy than Sad ($p < .001$), and vice versa for faces from the sadness chain ($p < .001$). These ratings, when being mapped onto the position of faces in the multidimensional space, show a graded change between categories (as an example of ratings on Happy, see Figure 2). Meanwhile, in chains explicitly optimizing ambiguity, ratings fluctuated non-linearly. Faces initially became more sad and less happy, but this trend reversed in later iterations, resulting in final faces that were rated as equally fitting Happy and Sad labels at a medium level.

These results confirmed that the Gibbs Sampling approach effectively optimized faces toward the intended direction (happy, sad, or ambiguous). The non-linear trajectory in the ambiguously happy chains might suggest that, at least in this condition, participants may have adopted different strategies at different points in the chain. One possibility is that: at the early stage of each optimization chain, participants adjusted the face to the direction of making it less probable to be labeled as happy. But after multiple cycles of iterations, the faces might become increasingly non-happy, which drives participants to adjust the faces to re-approach the category of happy.

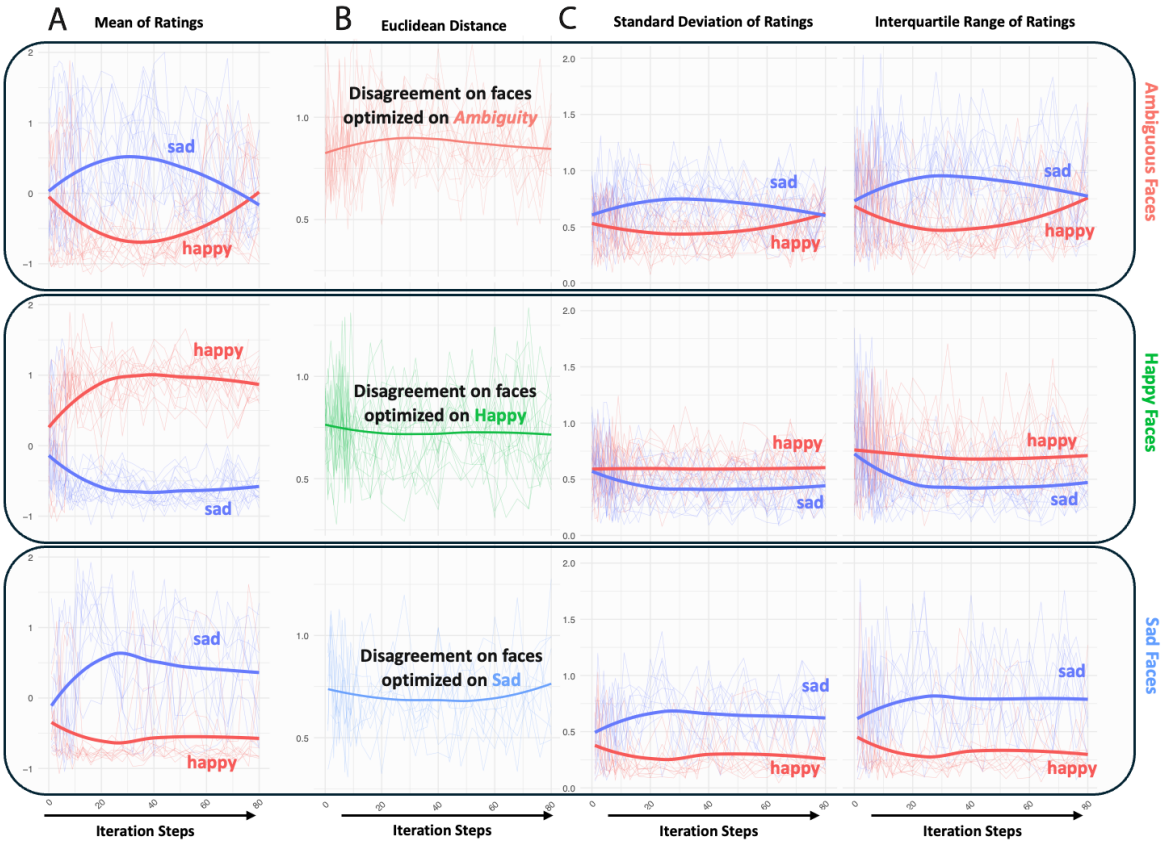

Figure 3: (A) The progression of mean Happy and Sad ratings on faces from each chain. (B) The progression of Euclidean distances showing disagreement on valence categorization. (C) The progression of standard deviation and interquartile range on Happy and Sad ratings. Thin lines in the background show progressions in each chains, and bold lines show the average across all chains.

Individual differences in interpreting valence. When optimizing ambiguity in matching the Happy label, the ambiguous faces were most likely optimized along a spectrum of ambivalence. According to the ambiguity account, ambiguously happy faces would be expected to elicit higher individual variation when categorized as either positive or negative. Although we did not directly collect ratings for "positive" and "negative" labels, we computed corresponding scores based on ratings along basic emotions that are valenced as positive or negative. As anticipated, for faces that have been optimized over at least one round, participants exhibited significantly greater disagreement in valence ratings for ambiguous faces ($M\ of\ distance = .88$, $SD = .17$) compared to faces from the happiness ($M\ of\ distance = .72$, $SD = .19$, $p < .001$) and sadness chains ($M\ of\ distance = .71$, $SD = .18$, $p < .001$; see Figure 3B).

Individual differences in labeling specific emotions. We next analyzed individual differences in applying specific emotion labels to faces across task conditions. When evaluating how well each face matches Happy, we found that: as predicted by a *conceptual diversity* account, we continued to find variation in Happy ratings of faces from the happiness chain after the initial cycle of optimization. There, the variation was even greater than that in the ambiguous condition (*SD*: $CI = [.07, .14]$, $p < .001$; *IQR*: $CI = [.10, .19]$, $p < .001$). In ratings of Sad, the individual variation in labeling consensually sad faces were only slightly lower than that in ambiguous faces ($CI = [-.10, -.02]$, $p = .001$, *IQR*: $CI = [-.16, -.04]$, $p < .001$, see Figure 3C). This observation challenges the intuitive expectation that prototypical happy and sad faces would be more consistently labeled than ambiguous ones. Instead, even for evidently categorized Happy faces, individuals still exhibit some degrees of disagreement.

One deflationary explanation of this finding is that people simply vary more when using 'high' values, such that higher ratings induce greater individual differences. We have addressed this concern in two ways. First, we have z-scored ratings within-participant to control for differences in use of the rating scale. Second, we modeled the standard deviation in ratings as a function of the mean ratings of corresponding faces, the emotion label that the rating is on (Happy vs. Sad), and the chain it came from (happiness vs. sadness), along with their interactions as fixed effects. The results revealed a

complex relationship between SD and mean ratings. Specifically, for faces optimized for sadness, a negative relation was found: as ratings on Sad increased, individual differences also increased ($\beta = .07$, $p = .003$). In contrast, for faces optimized for happiness, as ratings on Happy increased, variability decreased ($\beta = -.19$, $p < .001$). These findings suggest that variability in ratings cannot be solely explained by the nature of the distribution in using high ratings (IQR generally showed the same trend).

Still, individuals tended to show decreased variability when deciding how much the face matches an obviously inappropriate label (e.g., how sad a consensually happy face is, vice versa). It may be easier to agree on what a concept is *not* than what it is. Specifically, for faces optimized to be happy, individual ratings vary less on Sad ($M\ of\ SD = .46$, $SD = .22$) than Happy ratings ($M\ of\ SD = .59$, $SD = .20$); and vice versa for the consensually sad faces. This suggests that representations of emotions still generally align across individuals, having clear and shared boundaries defining what referents a semantic label unlikely refers to. It also rules out the possibility that the great individual differences in assigning an appropriate label are simply due to the noise in the continuous slider measurement.

Individual differences in labeling the ambiguous referents, on the other hand, fluctuated at a medium level. For example, when rating on Happy, ambiguous faces elicited less variation than faces optimized for happiness (see above), but more variation than those optimized for sadness (*SD*: $CI = [.16, .23]$, $p < .001$; *IQR*: $CI = [.19, .29]$, $p < .001$). Similar to the trend observed in the progression of averaged ratings, we observed that variability in Happy and Sad ratings on these faces first diverged but then converged. However, these changes are not statistically significant across iteration steps (*SD*: $\beta = .0006$, $p = .16$; *IQR*: $\beta = .0002$, $p = .73$). This may suggest that ambiguity elicits equivalent individual differences when being mapped onto all possible labels, as such referents fall within the gray intersection across well-structured categorical nodes.

General Discussion

In the present work, we identified two distinct sources of individual differences in semantic judgments: *referential ambiguity* and *conceptual diversity*. By leveraging Gibbs Sampling, we systematically generated facial stimuli that were either *ambiguous* or *prototypical* exemplars of basic emotion categories (Happy and Sad). Our findings reveal that ambiguous referents - faces that fell between categories - elicited significantly greater individual variation in valence judgments, consistent with the idea that less clear category membership drives divergence in information interpreting. Intriguingly, our results also show that even for faces that were initially chosen to be prototypical, there may be individual differences in how they are perceived by naive observers, suggesting that variability in conceptual representations extends beyond explicitly ambiguous cases. Taken together, we highlight the complex interplay between perceptual and conceptual factors in shaping individual differences in meaning-making, providing a framework for fine-grained distinguishing between these two paths to variation in semantic judgments.

One methodological concern is that Gibbs Sampling with People inherently samples from the population-level distribution, which may wash over the individual differences we aim to study. However, we note that this concern applies equally across all conditions: our findings demonstrate that even faces optimized for consensus exhibit meaningful individual variation from observers, supporting our claim that conceptual diversity extends beyond ambiguous cases. Future work must complement this approach with other methods that explicitly maximize individual differences to further explore the boundaries of semantic variation.

More broadly, our work contributes to understanding on how individual differences in semantic interpretation come into being. The two sources of individual differences we identified - referential ambiguity and conceptual diversity - align with distinct theoretical perspectives on concept acquisition. Referential ambiguity is predicted by the classical view that treats concepts as universal cognitive primitives existing prior to language (e.g. Fodor, 1975). Under this *nativist* view (e.g. Gleitman et al., 2005; Kuhl, 2000), language learning involves discovering how labels map onto pre-existing primitives in a language of thought, thus individual differences can be traced back to the noisiness or ambiguity in the surface mapping. However, this view predicts that concepts closer to direct sensory experiences – including our case of basic emotions – should be easier to learn, resulting in stronger alignment across individuals. This is inconsistent with our observation of variations in interpreting the confidently categorized faces, suggesting that the nature of semantic understanding may be more complex than solely retrospective mapping.

The diversity in conceptual representations is more readily explained by another perspective, which views labeling as an active procedure of categorizing the external world (e.g. Lupyan & Zettersten, 2021). Under this *constructionist* view, individual differences in semantic representations emerge from variations in personal experience, cultural background, and environmental context. Consequently, even seemingly clear-cut, well-defined concepts can exhibit substantial individual variability. Supporting this idea, cross-linguistic comparisons have shown that even concrete concepts, such as house and emotions, can differ significantly in meaning across cultural contexts (see Thompson et al., 2020).

Our study does not yet fully capture how these two mechanisms interact to produce the individual differences we observed. We also recognize the possible existence of other sources that may contribute to variability in semantic understanding. Identifying and disentangling these influences remains an important direction for future research. However, our work establishes a foundation for exploring these questions by demonstrating that individual differences in semantic interpretations can stem from multiple sources and can be systematically examined using appropriate tools.

Acknowledgement

We thank Lars Kotthoff for providing the d3 plugin we used to construct the space of Chernoff faces (https://github.com/gnarmis/chernoff-d3.git). We also thank Sonia Murthy for giving us insightful feedback on the project.

References

Botch, T. L., & Finn, E. S. (2024). Neural representations of concreteness and concrete concepts are specific to the individual. *Journal of Neuroscience*, *44*(45).

Brooks, J. A., & Freeman, J. B. (2018). Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature human behaviour*, *2*(8), 581–591.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American statistical Association*, *68*(342), 361–368.

Duan, Y., & Lupyan, G. (2023). Divergence in word meanings and its consequence for communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings* (Vol. 11). Elsevier.

Feldman, J., & Choi, L.-S. (2022). Meaning and reference from a probabilistic point of view. *Cognition*, *223*, 105058.

Fodor, J. (1975). *The language of thought*. Harvard University Press.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language learning and development*, *1*(1), 23–64.

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people. *Advances in neural information processing systems*, *33*, 10659–10671.

Johns, B. T. (2024). Determining the relativity of word meanings through the construction of individualized models of semantic memory. *Cognitive Science*, *48*(2), e13413.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–11857.

Li, P., Schloss, B., & Follmer, D. J. (2017). Speaking two "languages" in america: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behavior research methods*, *49*, 1668–1685.

Li, Z., Lu, H., Liu, D., Yu, A. N., & Gendron, M. (2023). Emotional event perception is related to lexical complexity and emotion knowledge. *Communications Psychology*, *1*(1), 45.

Lupyan, G., & Zettersten, M. (2021). Does vocabulary help structure the mind? *Minnesota symposia on child psychology: Human communication: Origins, mechanisms, and functions*, *40*, 160–199.

Marti, L., Wu, S., Piantadosi, S. T., & Kidd, C. (2023). Latent diversity in human concepts. *Open Mind*, *7*, 79–92.

Murthy, S. K., Griffiths, T. L., & Hawkins, R. D. (2022). Shades of confusion: Lexical uncertainty modulates ad hoc coordination in an interactive communication task. *Cognition*, *225*, 105152.

Nelson, N. L., & Russell, J. A. (2013). Universality revisited. *Emotion Review*, *5*(1), 8–15.

Oktar, K., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024). Dimensions of disagreement: Divergence and misalignment in cognitive science and artificial intelligence. *Decision*.

Quine, W. V. O. (2013). *Word and object*. MIT press.

Sarafoglou, A., Giacobello, A., Godmann, H., Johnson, T., Visser, I., Haaf, J. M., & Szymanik, J. (2024). A bayesian framework to study individual differences in semantic representations.

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behavior*, *4*, 1029–1038.

Van Geert, E., & Jacoby, N. (2024). Using gibbs sampling with people to characterize perceptual and aesthetic evaluations in multidimensional visual stimulus space. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Wang, X., & Bi, Y. (2021). Idiosyncratic tower of babel: Individual differences in word-meaning representation increase as word abstractness increases. *Psychological Science*, *32*(10), 1617–1635.

Winter, K. A., & Kuiper, N. A. (1997). Individual differences in the experience of emotions. *Clinical psychology review*, *17*(7), 791–821.