Gibbs Sampling with People

Abstract

A core problem in cognitive science and machine learning is to understand how humans derive semantic representations from perceptual objects, such as color from an apple, pleasantness from a musical chord, or trustworthiness from a face. Markov Chain Monte Carlo with People (MCMCP) is a prominent method for studying such representations, in which participants are presented with binary choice trials constructed such that the decisions follow a Markov Chain Monte Carlo acceptance rule. However, while MCMCP has strong asymptotic properties, its binary choice paradigm generates relatively little information per trial, and its local proposal function makes it slow to explore the parameter space and find the modes of the distribution. Here we therefore generalize MCMCP to a continuous-sampling paradigm, where in each iteration the participant uses a slider to continuously manipulate a single stimulus dimension to optimize a given criterion such as ‘pleasantness’. We formulate both methods from a utility-theory perspective, and show that the new method can be interpreted as ‘Gibbs Sampling with People’ (GSP). Further, we introduce an aggregation parameter to the transition step, and show that this parameter can be manipulated to flexibly shift between Gibbs sampling and deterministic optimization. In an initial study, we show GSP clearly outperforming MCMCP; we then show that GSP provides novel and interpretable results in three other domains, namely musical chords, vocal emotions, and faces. We validate these results through large-scale perceptual rating experiments. The final experiments combine GSP with a state-of-the-art image synthesis network (StyleGAN) and a recent network interpretability technique (GANSpace), enabling GSP to efficiently explore high-dimensional perceptual spaces, and demonstrating how GSP can be a powerful tool for jointly characterizing semantic representations in humans and machines.
1 Introduction

Humans continuously extract semantic representations from complex perceptual inputs, re-expressing them as meaningful information that can be efficiently communicated to primary cognitive processes such as memory, decision-making, and language [1–3]. Effective semantic representation seems to be a prerequisite for intelligent behavior, and is correspondingly a core topic of study in both cognitive science and machine learning [4–6].

One way of studying semantic representations in humans is to present participants with stimuli that exhaustively sample from a stimulus space (e.g., the space of visible colors) and ask them to evaluate these stimuli [e.g. 7]. Unfortunately, this method works poorly for high-dimensional stimuli whose parameter spaces are too large to explore exhaustively. An alternative approach is to hand-construct stimulus sets to test specific hypotheses about semantic representations [e.g., that happy faces look more trustworthy when facing the perceiver,8]; however, this approach relies heavily on prior domain knowledge, and is poorly suited to exploratory research.

Markov Chain Monte Carlo with People (MCMCP) addresses this problem [9–11]. This method presents participants with a choice between two stimuli (e.g., animals) and asks which comes from a given category (e.g., cats). Crucially, stimulus generation becomes progressively biased towards parts of the stimulus space that previously received positive category judgements, using a procedure inspired by the Markov Chain Monte Carlo (MCMC) technique from computational statistics. This approach has been shown to outperform reverse correlation, a related non-adaptive method for mapping semantic categories to perceptual spaces [12].

According to the underlying theory, MCMCP converges asymptotically to the participant’s internal probabilistic representation of a given semantic category [9]. However, its practical convergence rate is limited for several reasons. The first concerns the response interface: MCMCP is traditionally limited to binary choice responses, which can only provide a small amount of information per trial (1 bit), much less than the theoretical limit of other response interfaces (e.g., sliders). The second depends on the proposal function that generates successive stimuli: a too-narrow proposal function makes the process slow to find the modes of the distribution, whereas a too-wide proposal function makes it harder for the process to estimate the mode with much precision [13] [see also 14].

Here we present a new technique for addressing these problems, termed Gibbs Sampling with People (GSP). While MCMCP corresponds to a human instantiation of the Metropolis-Hastings MCMC sampler, GSP corresponds to a human instantiation of a Gibbs sampler. Crucially, unlike [15], our Gibbs sampler has participants respond with a continuous slider rather than a binary choice. This has two effects: first, it substantially increases the upper bound of information per trial, and second, it eliminates the need to calibrate a proposal function. We further show how GSP can be formulated in utility theory, thereby generalizing the approach from discrete to continuous semantic representations, and we show how GSP can be shifted towards deterministic optimization through an aggregation procedure.

A related paradigm with a Gibbs sampler interpretation is serial reproduction, where one participant’s imitation of a stimulus becomes the next stimulus in the chain [e.g. 17–22, 66]. However, serial reproduction is limited to percepts that can be entirely reproduced in a single trial [e.g., spoken sentences, 22]. In contrast, GSP participants only ever have to manipulate one stimulus dimension at a time, even if the stimulus itself is high-dimensional. This allows GSP to explore much richer stimulus spaces.

This paper continues with a review of MCMCP and a theoretical account of GSP. We then describe four studies applying GSP to various visual and auditory domains, ranging from simple low-dimensional problems to complex high-dimensional problems parameterized by deep neural networks. These studies include experiments directly implementing GSP and MCMCP, control experiments investigating different hyperparameters, and validation experiments for the generated outputs. All combined, these 16 experiments represent data from 2,960 human participants.
2 Theory

2.1 MCMC and MCMCP

MCMC is a procedure for sampling from distributions where the normalization constant is impractical to compute directly. It works by constructing a Markov chain whose stationary distribution corresponds to the probability distribution of interest; given enough samples from this Markov chain, it is then possible to approximate the probability distribution arbitrarily closely.

The MCMC algorithm may be performed by choosing an arbitrary initial state for the Markov chain, then repeating the following steps until convergence:

1. Sample a candidate $x^*$ for the next state of the Markov chain according to some proposal function $q(x^*; x)$.

2. Decide whether to accept this candidate according to an appropriate acceptance function $A(x^*; x)$ constructed to reflect the probability distribution of interest.

In the case of a symmetric proposal function, the acceptance function takes a simple form known as the Barker acceptance function and is given by $A(x^*; x) = \pi(x^*)/(\pi(x) + \pi(x^*))$ where $\pi(x)$ is the target distribution [23].

In MCMC the acceptance function is replaced with a human participant, whose task is to choose between the current state and the candidate state [9]. The trick is then to frame this task such that the participant’s choices correspond to the acceptance function for an identifiable probability distribution. The solution presented in the original MCMCP paper is to tell the participant that one of the stimuli comes from a class distribution (e.g., cats), and one of them comes from an unknown category. The authors suppose that the participant computes the posterior probability of class membership assuming a locally uniform likelihood for the alternative class, and then selects a stimulus with probability proportional to its posterior probability of class membership. Under these assumptions, the participant’s behaviour can be shown to correspond to the classic Barker acceptance function where the underlying probability distribution is the likelihood function for the class being judged.

This formulation is elegant but it has two important limitations. First, it can only be applied to semantic representations that take categorical forms; the derivation does not make sense for continuous semantic representations (e.g., attractiveness). Second, it assumes that participants make their choices with probabilities equal to their posterior probabilities of class membership (a process termed probability matching) as opposed to the Bayes-optimal strategy of deterministically maximizing this posterior probability [24]. Humans do indeed seem to exhibit probability matching in certain contexts, but a convincing cognitive model ought to explain how this sub-optimal process arises [25].

Here we reformulate MCMC (and later GSP) without these limitations. We suppose that the participant is asked to select the stimulus that best matches a given criterion $C$; for example ‘select the most attractive face’ or ‘select the color that most resembles lavender’. We suppose that the participant performs this task by extracting a utility value for each stimulus, and selecting the stimulus with the maximum utility [50]. In the case of the class membership tasks typically used in MCMC, we might hypothesize that the utility corresponds to the subjective likelihood of the stimulus conditioned on the class of interest; however, in the general case the utility function need not necessarily correspond to a probability distribution. The utility value is however assumed to have a deterministic component and a noise component, namely $U_i = \ell_i + n_i$, where $\ell_i$ is the deterministic utility of stimulus $i$, and $n_i$ is the associated noise. This noise component can capture participant-level noise from sensory [47] and cognitive [48] processes, as well as population-level noise corresponding to individual differences in the utility function [49]. In the case where the noise components are i.i.d. and have an extreme value distribution common to discrete choice models, it can be shown that the probability of selecting a given stimulus $s_1$ is equal to $(1 + \exp(-\gamma(\ell_1 - \ell_2)))^{-1}$ where $\gamma$ corresponds to the scale parameter of the noise component. If the utility is assigned based on subjective likelihood $\ell_i = \log p(s_i|C)$, then this equation reproduces the Barker acceptance function with target distribution $\pi(s) \propto p(s|C)$ . This justifies MCMC for an optimal observer with a noisy utility function (for proof and discussion see Appendix A.1).
2.2 Gibbs Sampling with People

Gibbs sampling is an alternative approach for sampling from probability distributions \[30\]. It is typically defined by the following algorithm: Let \( p(z_1, \ldots, z_n) \) be a target distribution over an \( n \)-dimensional state space from which one would like to sample. Then,

1. Choose a starting vector state \( z^{(1)} = (z_1^{(1)}, \ldots, z_n^{(1)}) \);
2. Circularly update coordinates by sampling from \( p(z_k^{(i+1)} | z_1^{(i+1)}, \ldots, z_{k-1}^{(i+1)}, z_{k+1}^{(i)}, \ldots, z_n^{(i)}) \).

In GSP we have the human participant provide the sample from the conditional distribution. This is achieved by presenting the participant with a slider that is associated with the current stimulus dimension \( z_k \) and instructing the participant to move the slider to maximize a certain criterion, such as the pleasantness of a sound or the resemblance of a fruit to a strawberry. To analyze the decision step, let us discretize the slider into a set of points \( \{ z_k^i \} \), and let \( z_{-k} \) denote the other fixed dimensions. As before, suppose that each point on the slider is associated with a utility that contains both a deterministic \( \ell(z_k^i) \) and a noise \( n_k \) component, and suppose that the participant chooses the slider location that maximizes the utility. Then, under similar assumptions to those made in the MCMCP case, it can be shown that the probability distribution over slider locations is given by

\[
p(\text{choose } i) = p(z_k^i | z_{-k}) = \frac{e^{\gamma \ell(z_k^i, z_{-k})}}{\sum_j e^{\gamma \ell(z_j^i, z_{-k})}} \tag{1}\]

and as the granularity of the slider tends to infinity, the denominator becomes a marginal, and GSP becomes a sampler from \( p(z) \propto e^{\gamma \ell(z)} \) (for proof and discussion see Appendix A.2).

A related method was used in [13], where participants imputed observations in sequences of coin flips in a manner that can also be interpreted as a Gibbs sampler. Our proposed approach differs in three important ways: a) our participants make continuous rather than binary choices, yielding significantly more information on each trial; b) we incorporate a notion of aggregation that continuously interpolates between sampling and optimization; c) we focus our method on understanding perceptual spaces, rather than on understanding subjective judgements of randomness.

As with MCMCP, GSP can be used to explore either categorical or continuous semantic representations. In the former case, the experimenter might ask a question like ‘adjust the slider until the image most resembles a dog’, and the participant’s utility function might correspond to the log probability of the image given the class, \( \ell(z) = \log p(z|C) \); in this case the sampler’s stationary distribution will be proportional to \( p^*(z|C) \). In continuous semantic representations, the utility function may not correspond to a probability distribution, and the interpretation would simply be that the sampler explores different regions of the space in proportion to their exponentiated utility \( e^{\gamma \ell(z)} \).

The noise parameter \( \gamma^{-1} \) is important for the behavior of the sampler. As \( \gamma^{-1} \to 0 \), the choice distribution becomes increasingly peaked around the highest utility item on the slider, shifting thus the sampler into an optimisation regime (specifically, coordinate descent). Typically we are interested in minimizing \( \gamma^{-1} \), so as to maximize the utility of the samples (mode seeking); however, some noise is still desirable because it helps drive exploration of the utility space.

There are two main ways to reduce the effective noise, \( \gamma^{-1} \). One approach is to estimate the joint distribution \( p(z) \propto e^{\gamma \ell(z)} \) by fitting a kernel density estimate (KDE) to the GSP samples, then simulating \( \gamma^{-1} \to 0 \) by taking the distribution’s mode. For simple distributions, this mode can also be estimated by averaging over samples. However, neither KDEs nor averaging are well-suited to complex high-dimensional spaces, where the joint distribution is hard to estimate reliably.

An alternative approach is to manipulate the Gibbs sampler itself. Specifically, suppose that we collect multiple samples from the conditional distribution of a given step of the Gibbs sampler \( p(\text{choose } i) \propto e^{\gamma \ell(z_k^i, z_{-k})} \), and then simulate \( \gamma^{-1} \to 0 \) by returning the peak of a KDE from these samples (or potentially the sample mean). This will in turn simulate \( \gamma^{-1} \to 0 \) for the joint distribution \( p(z) \propto e^{\gamma \ell(z)} \) as produced by the Gibbs sampler. The practical advantages of this approach are that a) we restrict density estimation to a more tractable one-dimensional case, and b) the same stimulus can be re-used for multiple trials, which can be useful when stimuli are slow to create. We explore both KDE and mean aggregation in this paper.
There are several ways to assign the iterations of a GSP or MCMCP chain to human participants. In within-participant chains, the entire chain is completed by just one participant, and the resulting samples reflect the semantic representations of that single participant. In across-participant chains, each iteration comes from a different participant, and the samples then reflect shared semantic representations across participants. We demonstrate both kinds of chains in this paper.

3 Studies

3.1 Color

Our first study concerns a particularly low-dimensional perceptual space: color. This kind of perceptual space should provide a useful sanity check for any semantic sampling procedure: if a procedure fails here, it is even less likely to succeed in high-dimensional perceptual spaces.

We compared three sampling methods for recovering the colors associated with eight words: ‘chocolate’, ‘cloud’, ‘eggshell’, ‘grass’, ‘lavender’, ‘lemon’, ‘strawberry’, and ‘sunset’. We parameterized color space using the perceptually oriented HSL scheme, where each color is encoded as three integers: hue, saturation, and lightness, taking values in [0, 360], [0, 100], and [0, 100] respectively.

The first sampling method was MCMCP, implemented with a Gaussian proposal function of standard deviation 30. The second method was standard GSP. The third method was aggregated GSP, collecting 10 slider responses for each iteration, and propagating the mean response to the next iteration. Each method was evaluated using across-participant chains of length 30, with five chains per color category (Experiments 1a, 1b, 1c). All participants (N = 422) were recruited from Amazon Mechanical Turk, and pre-screened with a color-blindness test and a color-vocabulary test before continuing with the online experiment. Each participant contributed up to 40 trials for a given sampling method. In each case, the participant was presented with a word (e.g., ‘lavender’), and asked to choose the color that best matched that word with either a binary choice interface (MCMCP) or a slider (GSP) (Figure 1A).

We then validated the resulting samples with a rating experiment, eliciting five ratings for each sample from a new participant group (N = 322, Experiment 1e). Participants were presented with the target word from the original chain, and asked to judge how well the color matched this word on a scale from 1 (not at all) to 4 (very much). The results indicate a clear advantage for GSP over MCMCP, with GSP converging faster and on higher ratings; they also show that aggregation robustly improved ratings (Figures 1B, 1C). Inspecting Figure 1B, it is evident that many of the MCMCP samples were poor reflections of their semantic category; meanwhile, GSP produced considerably fewer poor samples, and aggregated GSP even fewer. Investigating further, we found that the poor performance of MCMCP persisted when a) normalizing for the longer average duration of GSP trials, b) applying post-hoc aggregation, and c) trying different proposal widths (Experiment 1d, 153 participants. Appendix C). The implication is that, when the stimulus space is well-parameterized, GSP substantially improves sampling quality over MCMCP. In addition, it is clear that aggregation improves sampling quality still more at the cost of additional participant trials.

3.2 Emotional prosody

Our second study concerns a long-standing question in the psychological literature: how the way that a sentence is spoken (its prosody) communicates the emotional state of a speaker. Previous research into speech prosody mostly depends on analyzing recordings of actors expressing particular emotions, but such stereotypical recordings might not fully reflect ordinary emotion perception. GSP provides a way to study prosody perception without actors, instead generating emotional prosody directly from the perceptual judgments of listeners.

We began by taking three sentences from the Harvard sentence corpus recorded by a female speaker. These sentences are phonologically balanced and semantically neutral. We then defined our stimulus space in terms of seven acoustic manipulations that could be applied to these sentences to various degrees, corresponding to duration (speeding up or slowing down the fragment), intensity variation (rate and depth), and pitch (absolute level, range, slope, and F0 perturbation). We explored this stimulus space using 220 within-participant GSP chains, each comprising 21 iterations (Experiment 2a). Participants (N = 110) were recruited from Amazon Mechanical Turk.

Additional methods, results, and demographic information for all experiments are provided in the Appendix.
pre-screened with the audio test of [52], and instructed to manipulate the stimuli to reproduce one of three emotions: ‘anger’, ‘happiness’, and ‘sadness’ (Figure 2A). Each participant contributed two chains to the experiment.

Figure 2B plots mean feature values for the different emotional categories. Sad speech was marked by long duration, reduced pitch range, shallow pitch slope, and high F0 perturbation. Happiness had short duration, increased mean pitch, shallow pitch slope, and high pitch range. Anger had short duration, low mean pitch, falling pitch slope, and high pitch range. These characterizations are generally consistent with previous research [e.g., 61]. We also observed interesting patterns of feature correlations. For example, we found duration and F0 perturbation to be correlated for sadness ($r = .28$) but not for the other emotions (anger: $r = -.03$, happiness: $r = .00$); in contrast we found that pitch level and pitch slope were positively correlated in all three emotions (Appendix D.3). This suggests a new way to explore the perceptual spaces of perceived emotions, contrasting with previous literature that mainly focuses on unique contributions of single dimensions.

We then evaluated the resulting samples with a new participant group ($N = 161$), who rated how well samples matched each emotion on a four-point scale, producing c. 5.4 ratings per stimulus (Experiment 2c). Ratings increased steadily for the first sweep of the parameter vector and then plateaued with a reliable mean contrast of 0.9 points. We also replicated the results with across-instead of within-participant chains (Experiments 2b, 2d, 210 participants, Appendix D). These results imply that GSP is effective for exploring emotional prosody, and for generating emotional stimuli without the confounds of acted recordings.

3.3 Musical chords

Our third study concerns the subjective pleasantness of musical pitch combinations, or chords. For Westerners, this domain is highly multimodal, containing many prototypes of ‘pleasant’ (or ‘consonant’) chords. Exhaustively exploring this continuous space is difficult for conventional methods, and so far such research has been limited to single pitch intervals or to specific tuning systems [38–41]. Here we investigate whether GSP can help us characterize the continuous space of pairs of pitch intervals without restricting stimuli to a given tuning system.

Our stimulus space comprised two continuous intervals, specifying the logarithmic distance from the bass tone in the range 0.5–11. The standard Western tuning system corresponds to integer coordinates in this space. We explored this space with 50 across-participant GSP chains, each of length 40, created by 134 Mechanical Turk participants pre-screened with the audio test of [52] (Experiment 3a). These participants were instructed to make each chord as ‘pleasant’ as possible (Figure 3A). In a subsequent validation experiment, participants ($N = 168$) rated pleasantness for samples from a) the empirical distribution and b) the top five modes of KDEs applied to raw samples from iteration 10 onwards (Experiment 3b). Each condition received 662 ratings with up to 80 ratings per participant.
Ratings increased clearly as a function of iteration, with KDE modes scoring significantly higher than raw samples, supporting the utility of the aggregation method. The resulting KDEs display a rich structure that replicates and extends previous work in this area (Figure 3B) [38–41]. In particular, the 1D KDE shows clear integer peaks corresponding to the Western tuning system, with dips at the semitone (1) and tritone (6); the 2D KDE additionally shows peaks at various prototypical sonorities from Western music, such as the major triad (4, 7), the first inversion of the major triad (3, 8), and a dominant seventh chord with omitted third (7, 10) [see e.g., 42]. These results imply that GSP is an effective method for exploring continuous musical spaces.

3.4 Faces

Our final study addresses a particularly high-dimensional domain: images of human faces. Such images would be too high-dimensional to manipulate using GSP if they were expressed as raw vectors of pixel luminances, so we instead use a generative model to parameterize the perceptual space. State-of-the-art image synthesis models typically also have very high-dimensional parameter spaces, but here we build on recent work showing that the latent space of these models can be navigated surprisingly effectively using principal component analysis [PCA, 69]. Following [69], we apply this approach to the generative adversarial network ‘StyleGAN’ [67, 68], pretrained on the FFHQ dataset of faces from Flickr [67], and applying PCA to the intermediate latent code (termed \( w \) in the original papers). We used the top 10 PCA components to parameterize our stimulus space, allowing these components to vary up to two standard deviations from the mean, and fixing the input latent code (\( z \) in the original papers) to the population mean to control variability. We constructed 18 GSP chains for this stimulus space, with three chains of length 50 for each of the following adjectives, ‘attractive’, ‘fun’, ‘intelligent’, ‘serious’, ‘trustworthy’, and ‘youthful’ (Figure 4A, Experiment 4a), using 293 US participants from Mechanical Turk. We used an aggregation parameter of 5 trials per iteration, and used the arithmetic mean as the aggregation method. We then evaluated the generated samples with a behavioral rating experiment, following the same procedure as the color experiment but collecting 50 ratings per sample from 179 US participants (Experiment 4d).

The results are illustrated in Figure 4B–C. The GSP chains converged on good samples surprisingly quickly, with one full sweep of the 10 dimensions being sufficient to effectively capture the target categories as evaluated by the validation experiment. We also conducted two control experiments to investigate a) the effect of aggregating with KDE modes instead of arithmetic means (Experiment 4b,
4a, 539 participants) and b) the effect of widening to participant pool to include non-US participants (Experiments 4c, 4e, 530 participants, Appendix F). We found little difference from these manipulations, suggesting that multimodality was not a serious problem, and that the paradigm is robust to some cultural variation. Interestingly, the results hinted at certain relaxing of stereotypes in the non-US participants (e.g., producing intelligent samples that included women and non-Caucasian men). It would be worth exploring these variations in a systematic cross-cultural study.

4 Summary and conclusion

We have presented GSP, a new technique for extracting semantic representations from human participants. GSP organizes these participants into virtual Gibbs samplers, and thereby generates stimuli from the perceptual space associated with a given semantic representation. We have shown how this technique can recover semantic representations for a variety of perceptual domains, including color, emotional prosody, musical chords, and faces. The richness of the derived representations is compelling, and suggests many future psychological applications.

GSP has several features that seem to help it converge quickly on high-quality samples. One is its continuous-slider interface, which can deliver much more information per trial than the binary choice method used by MCMCP. A second is its lack of tuning parameter, which reduces the resources required to develop a workable experiment. A third is the way in which it manipulates a single stimulus dimension at a time: it is plausible that participants find it easier to evaluate differences between stimuli when the stimuli differ on just a single perceptual dimension.

By formulating GSP and MCMCP in utility theory [50] in contrast to MCMCP’s original Bayesian account [9], we enable both methods to be applied to continuous as well as categorical semantic representations, while relaxing assumptions about the participant’s prior and response noise. By incorporating aggregation into the conditional part of the Gibbs sampler, we increase the participant-to-stimulus ratio and thereby make GSP practical for stimuli that take a long time to generate, with the useful byproduct of averaging out perceptual noise. An interesting related possibility is to incorporate aggregation into MCMCP, producing something similar to simulated annealing [46].

The final study showed that GSP can work particularly well in combination with deep neural synthesis models. The important prerequisite is finding a relatively low-dimensional basis for the network for GSP to parameterize; fortunately, it seems that relatively simple techniques such as PCA can
sometimes suffice for this task [69]. The resulting combination provides a powerful link between cognitive science and deep learning, with GSP providing a way to jointly characterize human semantic representations in both perceptual space and the latent space of deep neural networks.

**Broader Impact**

This research extends the methods available to psychologists and cognitive scientists who seek to characterize semantic representations in human participants, and optionally link these representations to machine-learning models. In particular, the proposed method facilitates studying much richer perceptual spaces (both in terms of dimensionality and in terms of granularity) than can be explored effectively with conventional methods. Outside of fundamental research, these techniques could also be used in industry as way of compiling human feedback for complex models, and for developing software that helps users to manipulate complex deep-learning models to achieve practical or artistic goals. We can also see our methods helping also to improve human-computer interaction; for example, our speech prosody study is one step in the direction of improving the computer’s ability to understand and generate emotional speech.

When applying the GSP technique in combination with a machine-learning model, the generated results are influenced both by the participant’s underlying semantic representations and the biases inherent to the model and its training data. If not used carefully, the method could therefore potentially disadvantage populations who already suffer from various forms of discrimination (e.g., women, ethnic minorities). For example, in the face study we observed clear gender biases, with intelligence generally being associated with white men and attractiveness generally being associated with white women. These biases may be an arbitrary feature of the dataset (compiled from public Flickr accounts) as well as in the minds of the participants (recruited from Amazon Mechanical Turk), who bring their own culturally specific biases and experiences to bear on their judgments. Indeed, when we conducted a control experiment including non-US participants, we saw interesting hints at cultural divergences in facial stereotypes. These cultural issues deserve to be explored further.
Acknowledgments and Disclosure of Funding

The authors are grateful to David Poeppel for general help and support, and to Alec Mitchell, Jesse Snyder, Jordan Suchow, Matthew Wilkes, and Sally Kleinfeldt for their support of the Dallinger project.

References

[1] J. R. Anderson, *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.
[2] E. Rosch, “Cognitive representations of semantic categories,” *Journal of Experimental Psychology: General*, vol. 104, no. 3, pp. 192–233, 1975.
[3] M. N. Jones, J. Willits, S. Dennis, and M. Jones, *Models of semantic memory*, pp. 232–254. NY: Oxford University Press, 2015.
[4] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, “Topics in semantic representation,” *Psychological Review*, vol. 114, no. 2, pp. 211–244, 2007.
[5] G. Mesnil, A. Bordes, J. Weston, G. Chechik, and Y. Bengio, “Learning semantic representations of objects and their parts,” *Machine Learning*, vol. 94, no. 2, pp. 281–301, 2014.
[6] Q. Wang and K. Chen, “Alternative semantic representations for zero-shot human action recognition,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (New York, NY), pp. 87–102, Springer, 2017.
[7] P. Kay, B. Berlin, L. Maffi, W. R. Merrifield, and R. Cook, *The World Color Survey*. Stanford, CA: CSLI Publications, 2009.
[8] C. A. M. Sutherland, A. W. Young, and G. Rhodes, “Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties,” *British Journal of Psychology*, vol. 108, no. 2, pp. 397–415, 2017.
[9] A. Sanborn and T. L. Griffiths, “Markov Chain Monte Carlo with People,” in *Advances in Neural Information Processing Systems*, pp. 1265–1272, 2008.
[10] A. N. Sanborn, T. L. Griffiths, and R. M. Shiffrin, “Uncovering mental representations with Markov chain Monte Carlo,” *Cognitive Psychology*, vol. 60, no. 2, pp. 63–106, 2010.
[11] A. N. Sanborn and T. L. Griffiths, “Exploring the structure of mental representations by implementing computer algorithms with people,” in *Cognitive Modeling in Perception and Memory: A Festschrift for Richard M. Shiffrin*, pp. 212–228, Abingdon, UK: Taylor and Francis, 2015.
[12] J. B. Martin, T. L. Griffiths, and A. N. Sanborn, “Testing the efficiency of Markov chain Monte Carlo with people using facial affect categories,” *Cognitive Science*, vol. 36, no. 1, pp. 150–162, 2012.
[13] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds., *Markov Chain Monte Carlo in Practice*. London, UK: Chapman & Hall, 1996.
[14] A. Hsu, J. Martin, A. Sanborn, and T. Griffiths, “Identifying representations of categories of discrete items using Markov chain Monte Carlo with People,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2012.
[15] T. L. Griffiths, D. Daniels, J. L. Austerweil, and J. B. Tenenbaum, “Subjective randomness as statistical inference,” *Cognitive psychology*, vol. 103, pp. 85–109, 2018.
[16] N. Jacoby and J. H. McDermott, “Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction,” *Current Biology*, vol. 27, no. 3, pp. 359–370, 2017.
[17] V. Kempe, N. Gauvrit, and D. Forsyth, “Structure emerges faster during cultural transmission in children than in adults,” *Cognition*, vol. 136, pp. 247–254, 2015.
[18] J. Xu and T. L. Griffiths, “A rational analysis of the effects of memory biases on serial reproduction,” *Cognitive Psychology*, vol. 60, no. 2, pp. 107–126, 2010.
[19] S. Kirby, H. Cornish, and K. Smith, “Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 31, pp. 10681–10686, 2008.
[20] T. Verhoeven, S. Kirby, and B. De Boer, “Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals,” *Journal of Phonetics*, vol. 43, pp. 57–68, 2014.
[21] P. Edmiston, M. Perlman, and G. Lupyan, “Repeated imitation makes human vocalizations more word-like,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 285, no. 1874, 2018.
[22] B. Braun, G. Kochanski, E. Grabe, and B. S. Rosner, “Evidence for attractors in english intonation,” *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 4006–4015, 2006.
[23] A. A. Barker, “Monte Carlo calculations of the radial distribution functions for a proton-electron plasma,” *Australian Journal of Physics*, vol. 18, pp. 119–133, 1965.

[24] N. Vulkan, “An economist’s perspective on probability matching,” *Journal of Economic Surveys*, vol. 14, no. 1, pp. 101–118, 2000.

[25] D. R. Shanks, R. J. Tunney, and J. D. McCarthy, “A re-examination of probability matching and rational choice,” *Journal of Behavioral Decision Making*, vol. 15, no. 3, pp. 233–250, 2002.

[26] D. McFadden, “Conditional logit analysis of qualitative choice behaviour,” in *Frontiers in Econometrics* (P. Zarembka, ed.), pp. 105–142, New York, NY: Academic Press, 1974.

[27] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, “Motion illusions as optimal percepts,” *Nature Neuroscience*, vol. 5, no. 6, pp. 598–604, 2002.

[28] X.-X. Wei and A. A. Stocker, “A Bayesian observer model constrained by efficient coding can explain ‘anti-Bayesian’ percepts,” *Nature Neuroscience*, vol. 18, pp. 1509–1517, 2015.

[29] A. N. Sanborn, T. L. Griffiths, and D. J. Navarro, “A more rational model of categorization,” in *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (R. Sun and N. Miyake, eds.), pp. 726–731, Cognitive Science Society, 2006.

[30] A. E. Gelfand and A. F. M. Smith, “Sampling-based approaches to calculating marginal densities,” *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 398–409, 1990.

[31] G. H. Joblove and D. Greenberg, “Color spaces for computer graphics,” in *Proceedings of the 5th annual conference on Computer graphics and interactive techniques*, pp. 20–25, 1978.

[32] R. Banse and K. R. Scherer, “Acoustic profiles in vocal emotion expression,” *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[33] T. Bänziger, G. Hosoya, and K. R. Scherer, “Path models of vocal emotion communication,” *PLOS ONE*, vol. 10, Sept. 2015.

[34] “IEEE recommended practice for speech quality measurements,” tech. rep., IEEE, 1969. ISBN: 9781504402743.

[35] P. Demonte, “HARVARD corpus speech shaped noise and speech-modulated noise for SIN test,” 2019. Publisher: University of Salford.

[36] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, “Headphone screening to facilitate web-based auditory experiments,” *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.

[37] P. Laukka, H. A. Elfenbein, N. S. Thingujam, T. Rockstuhl, F. K. Iraki, W. Chui, and J. Althoff, “The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features,” *Journal of Personality and Social Psychology*, vol. 111, Nov. 2016.

[38] P. M. C. Harrison and M. T. Pearce, “Simultaneous consonance in music perception and composition,” *Psychological Review*, vol. 127, pp. 216–244, 2020.

[39] I. Lahdelma and T. Eerola, “Cultural familiarity and musical expertise impact the pleasantness of consonance/dissonance but not its perceived tension,” *Scientific Reports*, vol. 10, 2020.

[40] R. Plomp and W. J. M. Levelt, “Tonal consonance and critical bandwidth,” *The Journal of the Acoustical Society of America*, vol. 38, pp. 548–560, 1965.

[41] W. A. Sethares, *Tuning, timbre, spectrum, scale*. London, UK: Springer, 2005.

[42] D. L. Bowling, D. Purves, and K. Z. Gill, “Vocal similarity predicts the relative attraction of musical chords,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 1, pp. 216–221, 2018.

[43] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, “GANSpace: Discovering interpretable GAN controls,” *arXiv*, 2020.

[44] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pp. 4401–4410, 2019.

[45] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” *arXiv*, 2019.

[46] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[47] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, “Motion illusions as optimal percepts,” *Nature Neuroscience*, vol. 5, no. 6, pp. 598–604, 2002.

[48] X.-X. Wei and A. A. Stocker, “A Bayesian observer model constrained by efficient coding can explain ‘anti-Bayesian’ percepts,” *Nature Neuroscience*, vol. 18, pp. 1509–1517, 2015.
[49] A. N. Sanborn, T. L. Griffiths, and D. J. Navarro, “A more rational model of categorization,” in Proceedings of the 28th Annual Conference of the Cognitive Science Society (R. Sun and N. Miyake, eds.), pp. 726–731, Cognitive Science Society, 2006.

[50] D. McFadden, “Conditional logit analysis of qualitative choice behaviour,” in Frontiers in Econometrics (P. Zarembka, ed.), pp. 105–142, New York, NY: Academic Press, 1974.

[51] K. E. Train, Discrete choice methods with simulation. Cambridge University Press, 2009.

[52] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, “Headphone screening to facilitate web-based auditory experiments,” Attention, Perception, & Psychophysics, vol. 79, no. 7, pp. 2064–2072, 2017.

[53] J. H. Clark, “The Ishihara Test for color blindness,” American Journal of Physiological Optics, vol. 5, pp. 269–276, 1924.

[54] “IEEE recommended practice for speech quality measurements,” tech. rep., IEEE, 1969. ISBN: 9781504402743.

[55] P. Demonte, “HARVARD corpus speech shaped noise and speech-modulated noise for SIN test,” 2019. Publisher: University of Salford.

[56] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program].” Version 6.0.37, retrieved 3 February 2018. http://www.praat.org/.

[57] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” Journal of Phonetics, vol. 71, pp. 1–15, 2018.

[58] I. R. Titze, Y. Horii, and R. C. Scherer, “Some technical considerations in voice perturbation measurements,” Journal of Speech, Language, and Hearing Research, vol. 30, pp. 252–260, June 1987.

[59] R. Corretge, “Praat vocal toolkit.” http://www.praatvocaltoolkit.com, 2020.

[60] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner, “The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures,” Nature Human Behaviour, vol. 3, pp. 369–382, Apr. 2019.

[61] P. Laukka, H. A. Elfenbein, N. S. Thingujam, T. Rockstuhl, F. K. Iraki, W. Chui, and J. Althoff, “The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features,” Journal of Personality and Social Psychology, vol. 111, Nov. 2016.

[62] P. M. C. Harrison and M. T. Pearce, “Representing harmony in computational music cognition,” PsyArXiv, 2020.

[63] T. Stainsby and I. Cross, “The perception of pitch,” in The Oxford handbook of music psychology (S. Hallam, I. Cross, and M. Thaut, eds.), pp. 47–58, New York, NY: Oxford University Press, 2009.

[64] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” Science, vol. 344, no. 6191, pp. 1492–1496, 2014.

[65] X.-F. Wang and Y. Xu, “Fast clustering using adaptive density peak detection,” Statistical methods in medical research, vol. 26, no. 6, pp. 2800–2811, 2017.

[66] N. Jacoby and J. H. McDermott, “Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction,” Current Biology, vol. 27, no. 3, pp. 359–370, 2017.

[67] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR, pp. 4401–4410, 2019.

[68] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” arXiv, 2019.

[69] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, “GANSpace: Discovering interpretable GAN controls,” arXiv, 2020.
Appendix A  Mathematical framework

In this section we present derivations of the formulas derived in the theoretical exposition of the main text. We start from a derivation of the acceptance function of MCMCP based on utility theory which we then generalize to GSP.

A.1 MCMCP

To analyze the decision step of MCMCP, let us imagine that a participant is presented with two alternatives, $s_1$ and $s_2$, from which they are asked to choose according to some criterion $c$. In the context of utility theory, we suppose that the participant performs this task by extracting a utility value for each stimulus, and selecting the stimulus with the maximum utility. The utility value has two components, a deterministic component and a noise component, namely $U_i = \ell_i + n_i$, where $\ell_i$ is the deterministic utility of stimulus $i$, and $n_i$ is the associated noise. This noise component can capture participant-level noise from sensory [1, 2] and cognitive [2, 3] processes, as well as population-level noise corresponding to individual differences in the utility function [4]. In the current derivation, we assume that the noises are i.i.d. and that they are Gumbel distributed (also known as type I extreme value), $n_i \sim \text{Gumbel}(\mu, \gamma^{-1})$. Gumbel distributions are commonly used in discrete choice models because they approximate normal noise while possessing useful analytic properties [5]. From here, the probability of choosing, say $s_1$, would be

$$p(\text{choose } s_1) = p(U_1 > U_2) = p(n_2 - n_1 < \ell_1 - \ell_2)$$

(2)

To proceed, we recall the following useful property of Gumbel distributions: let $X \sim \text{Gumbel}(\mu, \gamma^{-1})$ and $Y \sim \text{Gumbel}(\mu_2, \gamma^{-1})$ be two independent variables, then the difference is logistically distributed, namely, $X - Y \sim \text{Logistic}(\mu_1 - \mu_2, \gamma^{-1})$. Thus, we see that the right hand side of (2) is simply the cumulative distribution function of the logistic distribution Logistic(0, $\gamma^{-1}$), from which it follows that

$$p(\text{choose } s_1) = \frac{1}{1 + e^{-\gamma(\ell_1 - \ell_2)}}$$

(3)

which is the desired result (see Section 2.1 in the main paper).

A.2 GSP

Let us now generalize the analysis of MCMCP to GSP. Recall that in the GSP step, a participant is presented with a slider that is associated with an active dimension, say $z_k$, from which they are asked to select a value. To analyze the decision step, let us discretize the slider into a set of points $\{z^+_k\}$, and let $z_{-k}$ denote the other fixed dimensions. Similar to the MCMCP case, we assume that the participant extracts a utility value for each stimulus along the slider, namely, $U_i = \ell(z^+_k, z_{-k}) + n_i$, with the noise being i.i.d. and Gumbel distributed $n_i \sim \text{Gumbel}(\mu, \gamma^{-1})$, and we assume that they choose the alternative with the highest utility. Such a choice model is known in the literature as the multinomial logit [5]. For completeness, let us derive the formula for the probability of choosing the alternative $z^+_k$. We have

$$p(z^+_k | z_{-k}) = p \left( \bigcap_{j \neq i} U_i > U_j \right)$$

$$= \int_0^\infty p(n_i > \ell_i - \ell_j + n_j | n_i = \epsilon) p(n_i = \epsilon)$$

$$= \int_0^\infty \prod_{j \neq i} p(n_j < \ell_i - \ell_j + \epsilon) p(n_i = \epsilon)$$

$$= \gamma \int_0^\infty d\epsilon \exp \left\{ - \left( 1 + \sum_{j \neq i} e^{-\gamma(\ell_i - \ell_j)} \right) e^{-\gamma(\epsilon - \mu)} - \gamma(\epsilon - \mu) \right\}$$

$$= \frac{e^{\gamma \ell_i}}{\sum_j e^{\gamma \ell_j}}$$
where in the third equality we used the fact that the noises are independent, and in the fourth equality we plugged in the standard formulas for the probability and cumulative distributions of the Gumbel distribution. The fifth equality follows from substituting $e' = \gamma(e - \mu)$ and noticing that the sum over exponentiated utility differences is a positive number, so that the integral identity $\int_{-\infty}^{\infty} dx \exp \left\{ -\lambda e^{-x} - x \right\} = 1/\lambda$ holds. Thus, substituting $\ell_i = \ell(z_k, z_{-k})$ we arrive at the desired equation, i.e. Equation (1) in the main paper. Notice also that in the case of two alternatives, this derivation recovers the acceptance function of MCMCP.

Both derivations of the MCMCP and GSP choice probabilities relied on two main assumptions regarding the nature of the noise: a) it is i.i.d., and b) it is Gumbel distributed. Starting from the latter, notice that the derivation of the GSP choice probability makes it clear how to generalize to other types of noise. Indeed, up to (and including) the third equality, we relied only on the i.i.d. nature of the noise. Moreover, the third equality provides a prescription on how to generalize: for a given choice of noise model, simply plug in the right cumulative function and probability distribution of that model. Thus, for a Gaussian noise for example, i.e. $n_i \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$p(z_k | z_{-k}) = \frac{1}{\sqrt{2\pi} \sigma^2} \int_{-\infty}^{\infty} d\epsilon \prod_{j \neq i} \Phi \left( \frac{\ell_i - \ell_j + \epsilon}{\sigma} \right) e^{-\frac{\epsilon^2}{2\sigma^2}}$$

where $\Phi$ is the normal cumulative distribution function. This is known as the independent probit model [5]. Of course, unlike the Gumbel case, in the Gaussian case this does not result in a closed form formula. This, however, does not prevent the GSP process from exploring the utility terrain of the model, given the functional similarity between Gumbel and normal distributions.

If the noise cannot be assumed to be i.i.d., the third equality no longer holds. We see two main potential sources of i.i.d. violations in this paradigm:

1. Intra-participant correlation (different participants have different utility functions);
2. Neighboring-point correlation (neighboring points on the slider are likely to receive correlated noise).

Intra-participant correlation has different implications for across- and within-participant chains (Figure 5). In across-participant chains, each participant only contributes one observation to the chain, so intra-participant correlation never manifests. In within-participant chains, all observations in a given chain come from the same participant, meaning that the i.i.d. assumption remains unviolated, and each chain ends up approximating the underlying utility function for the individual participant. The population-level utility function can then be approximated by aggregating over chains.

Neighboring-point correlation could have a subtle effect on the derivations presented here. Future work could revise our model to include a correlation structure for these points, for example following the correlated multinomial probit model where noise values are taken from a joint Gaussian distribution with a specified correlation structure [5].

Our derivation also does not take into account context effects, whereby the participant’s previous trials influence their responses to the present trial. In particular, it is possible that the utility value
for a given stimulus changes when the participant has already experienced that stimulus multiple times. This possibility is particularly high in within-participant chains, where the same participant experiences many stimuli from adjacent steps in the Gibbs sampler; in contrast, across-participant chains mostly avoid this effect by preventing the participant from experiencing multiple stimuli from the same chain (Figure 5). We tested the strength of this effect in the emotional prosody experiment, conducting both a within-participants and an across-participants version of the same paradigm. We found that the results did not differ materially between the two, implying that memory effects were not a significant problem for this paradigm.

The above GSP derivation also assumes that the participant visits all slider positions. If this assumption is violated, the denominator in the GSP choice probability would cover only a subset of the locations, effectively reducing the granularity of the slider. We can try and minimize this effect experimentally, by forcing the participant to explore a certain amount of the stimulus space before proceeding to the next trial, but it is often impractical to enforce exhaustive exploration. However, we expect that the consequences of this assumption violation are not severe for two reasons: a) participants tend to focus on the parts of the slider that contain most of the utility/probability mass, and b) participants can extrapolate between slider locations to estimate the utility values of intermediate points. Nonetheless, we would like to explore this assumption more in future work.

Appendix B General methods

B.1 Implementation

We implemented all experiments in PsyNet, our under-development framework for implementing complex experiment paradigms such as GSP and MCMCP. This framework builds on the Dallinger platform for experiment hosting and deployment. Participants interact with the experiment via their web browser, which communicates with a back-end Python server cluster responsible for organizing the timeline of the experiment (Figure 6). This cluster is mostly managed by Heroku and comprises a customizable collection of virtual instances that share the experiment management and stimulus generation workload, as well as a Postgres database instance for storing results. In some experiments we additionally used Amazon Web Services (AWS) S3 storage for hosting stimuli, and an AWS Elastic Compute Cloud (EC2) instance with an NVIDIA K80 GPU for deep neural network synthesis.

---

https://github.com/Dallinger/Dallinger
https://www.heroku.com/home
https://aws.amazon.com/
B.2 Participants

All participants provided informed consent in accordance with the Max Planck Society Ethics Council approved protocol (application 2018-38). All participants were recruited from Amazon Mechanical Turk, which is an online service for crowd-sourcing workers for online tasks. Here the only universal constraint we placed on recruitment was that participants must have a 95% or higher approval rate on previous tasks on Mechanical Turk; this approval criterion is meant to help recruit reliable participants. In some experiments we also constrained the worker to be a US resident.

When designing the experiment, each component was given an estimate for the average time it should take to complete; participants were then paid at a US $9/hour rate according to how much of the experiment they completed. Importantly, participants were still paid a proportional amount even if they left the experiment early on account of failing a pre-screening task.

A total of 2,960 participants took part in the 16 experiments reported in this paper, excluding those who failed pre-screening tests. For the participants who reported demographic information, ages ranged from 18 to 77 (\(M = 34.61, SD = 10.17\)), and 36.52% identified as female (63.12% male and 0.36% other).

These participants may be further differentiated into two groups: those who participated in the main experiments and those who participated in the validation experiments. These two groups had similar compositions, with the main participant group (Table 1) comprising 1,716 participants (35.31% female, 64.22% male, 0.47% other; ages 18–77, \(M = 34.33, SD = 9.86\)), and the validation group comprising 1,244 participants (38.35% female, 61.45% male, 0.21% other; ages 18–74, \(M = 35.03, SD = 10.60\)).

The musical chord study also collected additional information about musical expertise. Participants in the main chord experiment reported 0–25 (\(Med = 2, M = 4.26, SD = 6.23\)) years of musical experience (i.e., playing an instrument or singing), whereas participants in the corresponding validation experiment reported 0–64 (\(Med = 2, M = 4.39, SD = 7.57\)) years of musical experience.

Participant recruitment was managed by PsyNet. For the across-participant chain experiments, we specified a desired number of chains and a desired length for these chains, and participants were then automatically recruited until the chains reached their desired lengths. For the within-participant chain experiments, we specified a desired number of completed participant sessions, and recruitment continued until this threshold was met. For the rating experiments, we chose a desired number of ratings per experimental condition\(^7\) such that we expected that any variation in the resulting condition means should primarily reflect the stochasticity of the original sampler rather than the stochasticity of participant raters. As a rule of thumb, we aimed for approximately 150 participants per validation study, scaling this number accordingly when the validation study compared multiple methods. Participants were then automatically recruited until the minimum number of ratings per experimental condition was reached.

---

\(^5\)https://www.mturk.com

\(^6\)These demographics exclude the validation experiments for emotional prosody, for which demographic information was not collected.

\(^7\)An experimental condition typically corresponded to one point on a figure, for example iteration 10 for the ‘attractiveness’ GSP chains.
### Table 1: Main experiments

| Experiment | Method               | Rep. | Dim. | Iter. | Agg. | Chain type | N   | Pre-screening | US-only | Validated |
|------------|----------------------|------|------|-------|------|-------------|-----|---------------|---------|-----------|
| 1a Color   | MCMCP                | 8    | 3    | 30    | 1    | Across      | 57  | CB, CV        | No      | Exp. 1e   |
| 1b Color   | GSP                  | 8    | 3    | 30    | 1    | Across      | 53  | CB, CV        | No      | Exp. 1e   |
| 1c Color   | GSP (aggregation)    | 8    | 3    | 30    | 10   | Across      | 312 | CB, CV        | No      | Exp. 1e   |
| 1d Color   | MCMCP (proposal)     | 8    | 3    | 30    | 1    | Across      | 153 | CB, CV        | No      | -         |
| 2a Prosody | GSP (within)         | 3    | 7    | 21    | 1    | Within      | 110 | Audio         | Yes     | Exp. 2c, 2d|
| 2b Prosody | GSP (across)         | 3    | 7    | 20    | 1    | Across      | 57  | Audio         | Yes     | Exp. 2d   |
| 3a Musical chords | GSP     | 1    | 2    | 40    | 1    | Across      | 134 | Audio         | No      | Exp. 3b   |
| 4a Faces US| GSP (mean)           | 6    | 10   | 50    | 5    | Across      | 293 | CV            | Yes     | Exp. 4d, 4e|
| 4b Faces US| GSP (KDE)            | 6    | 10   | 50    | 5    | Across      | 278 | CV            | Yes     | Exp. 4e   |
| 4c Faces global | GSP (KDE)     | 6    | 10   | 50    | 5    | Across      | 269 | CV            | No      | Exp. 4e   |

**Note.** ‘Rep.’ indicates the number of semantic representations that were tested; ‘Dim.’ indicates the dimensionality of the stimulus space; ‘Iter.’ indicates the number of iterations in each chain; ‘Agg.’ indicates how many participants contributed to each iteration of the GSP chain; ‘N’ denotes the number of participants included in the final analysis; ‘CB’ denotes the color blindness pre-screening task; ‘CV’ denotes the color vocabulary pre-screening task; ‘US-only’ indicates whether the participant group was restricted to US residents; ‘Exp.’ denotes ‘Experiment’.

### Table 2: Validation experiments

| Experiment       | Ratings per participant | Ratings per stimulus | Total stimuli | N   | Pre-screening | US-only | Validating |
|------------------|-------------------------|----------------------|---------------|-----|---------------|---------|------------|
| 1e Color         | 60                      | 5.2                  | 3,720         | 322 | CB, CV        | No      | Exp. 1a, 1b, 1c |
| 2c Prosody       | 147                     | 5.4                  | 4,383         | 161 | Audio         | Yes     | Exp. 2a    |
| 2d Prosody       | 132                     | 4.1                  | 4,874         | 153 | Audio         | Yes     | Exp. 2a, 2b|
| 3b Musical chords | 80                      | 16.4                 | 820           | 168 | Audio         | Yes     | Exp. 3a    |
| 4d Faces         | 80                      | 52.1                 | 275           | 179 | CV            | Yes     | Exp. 4a    |
| 4e Faces         | 80                      | 25.6                 | 815           | 261 | CV            | No      | Exp. 4a, 4b, 4c|

**Note.** ‘N’ denotes the number of participants included in the analysis; ‘CB’ denotes the color blindness pre-screening task; ‘CV’ denotes the color vocabulary pre-screening task; ‘US-only’ indicates whether the participant group was restricted to US residents; ‘Exp.’ denotes ‘Experiment’.
B.3 Pre-screening tests

A useful technique for improving the quality of data from online participants is to implement pre-screening tests designed to screen out participants likely to deliver low-quality data [6]. Here we used three pre-screening tests in various combinations: a color blindness test, a color vocabulary test, and an audio test.

The color blindness test was derived from the well-known Ishihara color blindness test [7]. Here participants had to respond to six trials where the task was to transcribe a number from an image, with the contrast of the image being designed such that it is difficult to perform if the participant suffers from color perception deficiencies (see Figure 7 for an example trial). The image was set to disappear after three seconds to encourage quick responses. Each participant had to take six such trials; to pass, they had to answer at least four of these six trials correctly.

The color vocabulary test was constructed by taking six English color words that require a relatively good vocabulary knowledge to understand: ‘turquoise’, ‘magenta’, ‘granite’, ‘ivory’, ‘maroon’, and ‘navy’. None of these words were used in the other experiments. We associated each word with an RGB definition sourced from Wikipedia, and presented the participant with six trials where they were presented with a color and had to choose which of the eight words corresponded to that color (see Figure 8). The pass threshold was a score of four out of six.

The audio pre-screening task, originally developed in [6], was intended to ensure that participants were wearing headphones and could hear perceive subtle sound differences. The task has participants perform a three-alternative forced-choice task to identify the quietest of three tones. These tones are constructed to elicit a phase cancellation effect, such that when played on loudspeakers the order of quietness changes, causing the participant to respond incorrectly. Each participant had to take six such trials; to pass, they had to answer at least four of these six trials correctly.

B.4 Performance incentives

In order to further improve data quality, some of our experiments (specifically, all but the emotional prosody experiments) additionally included a financial incentive for participants to provide high-quality data. Prior to the main part of the experiment, we informed all participants of this incentive, using the following text:

The quality of your responses will be automatically monitored, and you will receive a bonus at the end of the experiment in proportion to your quality score. The best way to achieve a high score is to concentrate and give each trial your best attempt.
We purposefully left the definition of ‘quality’ vague, so as to avoid encouraging participants to ‘game’ a particular aspect of response quality. Of course, our tasks were subjective, and so there was no meaningful way to define a high-quality answer \textit{a priori}. Instead, our approach was to use consistency as a proxy for quality; the rationale is that a participant who takes the task seriously and carefully is likely to deliver consistent responses when administered the same trial multiple times, in contrast to a participant who does not pay attention to the task and simply answers randomly.

We estimated consistency as follows. Once a participant finished all of their ‘main’ experiment trials, they then received a small number (4–8, depending on the experiment) of trials that repeated randomly selected trials from the earlier part of the experiment. The data from these trials contributed solely to consistency estimation, not to chain construction. In GSP trials and four-point rating trials, consistency was quantified by taking the Spearman correlation between the two sets of answers; for MCMCP trials, consistency was quantified by taking the percentage agreement between the two sets of answers. Participants were then given a small monetary bonus in proportion to the resulting consistency score, ranging from zero dollars for chance performance up to one dollar for perfectly consistent performance.

\subsection*{B.5 Chain construction}

In all experiments except the emotional prosody experiment, we randomized the starting locations of each chain by randomly sampling from a uniform distribution over the range of permissible feature values. In the case of emotional prosody, we found this randomization problematic because it often led to unrealistic parts of the stimulus space. In this case we therefore initialized each chain at a ‘null’ state corresponding to the unaltered reference sentence.

In a given trial of a GSP experiment, the participant’s slider manipulated exactly one dimension of the stimulus. To counteract any potential biases towards left or right slider directions, we randomized the effective direction of the slider on each trial, such that approximately half of the time the right of
the slider corresponded to positive feature values, and the other half of the time it corresponded to negative feature values.

Our experiments implemented both within-participant and across-participant chains (Figure 5). In within-participant chains, the entire chain is completed by just one participant, and the resulting samples reflect the semantic representations of that single participant. In across-participant chains, each iteration comes from a different participant, and the samples then reflect shared semantic representations across participants.

Across-participant chains are more complex to implement because of the interaction between multiple participants. Each time a participant is ready to take a new trial, it is necessary to scan the different chains in the experiment and identify one that satisfies the following conditions:

1. The chain is not full (i.e., it has not reached its specified quota of iterations);
2. The participant has not already participated in that chain;
3. No other participants have been assigned to that particular iteration of the chain.

The last point – ensuring that multiple participants are not assigned to the same iteration of a chain – is important for the efficiency of data collection, but it can cause problems when a participant claims a particular iteration of the chain and then drops out of the experiment, potentially blocking any future additions to that chain. We therefore implemented a time-out parameter for this experiment, set to 60 seconds, after which the participant’s pending trial is invalidated and the chain is unblocked.

In within-participant chains we are free to discard all of a participant’s data when they drop out of an experiment partway through. This is not practical in across-participant chains, however, where many subsequent participants might have built on the data previously contributed by this participant. In the latter case, we therefore retain the participant’s contributions even when they drop out of the experiment.

Appendix C  Color

C.1 Supplementary methods

We chose eight words designed to be moderately but not overly familiar to English speakers that we anticipated to evoke strong color associations. These words were ‘chocolate’, ‘cloud’, ‘eggshell’, ‘grass’, ‘lavender’, ‘lemon’, ‘strawberry’, and ‘sunset’. We then explored the perceptual spaces associated with these eight words using GSP and MCMCP.

We implemented GSP and MCMCP using the Hue, Saturation, Lightness (HSL) color space. We chose this color space over the Red, Green, Blue (RGB) color space because it is generally considered to better reflect how humans perceive color relationships. In this space, each color is encoded as three integers: hue, saturation, and lightness, taking values in $[0, 360)$, $[0, 100]$, and $[0, 100]$ respectively.

MCMCP relies on the specification of a proposal function. In our main experiment, we used a Gaussian distribution with a standard deviation of 30, chosen on the basis of internal piloting, and rounding samples to the nearest integer values. In MCMCP the proposal distribution should be symmetric, which can be problematic to satisfy when the sampler reaches the boundaries of the sample space. We addressed this problem by computing the proposal distribution modulo the scale range, such that moving past the top of the scale means returning to the bottom of the scale. This works particularly well for hue, which is already defined as a circular space, with both 0 and 360 corresponding to the color red. It works less well for saturation and lightness, because these are linear scales; however, as we chose target colors occupying central regions of these two scales, we expected that these boundary effects should not materially influence MCMCP’s performance.

We implemented simple web interfaces for the MCMCP and GSP tasks. In the MCMCP task, participants were presented with pairs of colors, and had to choose which color best represented a target word (Figure 10). In the GSP task, participants were presented with a single color that constantly updated to reflect the current position of a slider; participants were then instructed to move the slider to make the color represent a target word as well as possible (Figure 11).

For each given sampling method we constructed five across-participants chains per adjective, yielding 40 chains in total. Each chain was filled to a length of 30 states, not including the initial random
Each participant contributed a maximum of 40 trials to the chains for a given sampling method (Experiments 1a–c; see Table 1 for participant numbers).

The aggregated GSP experiment combined 10 trials for each iteration of the Gibbs sampler (Experiment 1c). These 10 trials were combined using the arithmetic mean in the case of saturation and lightness, and the circular mean in the case of hue. These means were then propagated to the next iteration of the Gibbs sampler.

We also ran a control experiment where we tested the sensitivity of MCMCP to the width of its proposal function (Experiment 1d). We tested five different standard deviations for this proposal function: 10, 20, 30, 40, and 50, all expressed on the integer color scale.

The validation experiment (Experiment 1e) used the same pre-screening procedure as the chain-construction experiments (Experiments 1a–d). A minimum of five ratings were collected for each sample generated in the former experiments, with the constraint that participants could not rate the same stimulus more than once. Participants were assigned pseudo-randomly to stimuli such that the number of ratings accumulated evenly for each stimulus. In each trial, participants were presented with the target word from the original chain, and asked to judge how well the colour matched this word on a scale from 1 (not at all) to 4 (very much). A given participant’s ratings were only included in the final tallies if they completed the entire validation experiment. See Table 2 for participant numbers.

C.2 Supplementary results

In the main paper we identified a clear advantage for GSP over MCMCP, given chains of the same length and the same amount of aggregation. However, we were concerned about three possible confounds:

1. GSP trials are more time-consuming than MCMCP trials. Even if GSP requires fewer trials to achieve good sample quality, if these trials take much longer, then GSP will end up being practically slower than MCMCP.

2. A common analysis approach with MCMCP is to generate category prototypes by averaging over many samples. Perhaps MCMCP performs better when using this analysis method.

3. MCMCP has a tuning parameter corresponding to the width of the proposal function. Perhaps the relatively poor performance of MCMCP was simply due to the wrong choice of proposal width.

We will now describe how we addressed these potential confounds. Regarding the first issue, Figure 12 plots validation ratings for the different sampling methods, with the horizontal axis now corresponding to the total participant time invested in the respective chains (Experiments 1a–c). We estimated total participant time by taking iteration number and multiplying it by the median participant time spent on
the two different trial types. The results indicate that non-aggregated GSP still clearly outperformed MCMCP despite the longer duration of its individual trials. It is difficult to make a clear statement about the relative performance of aggregated GSP because its profile overlaps minimally with the other two methods; however, the figure implies that non-aggregated GSP outperforms aggregated GSP for the first few iterations, with aggregated GSP then overtaking at a later point. This is consistent with our expectations: the fast-but-noisy non-aggregated GSP can quickly escape its low-probability starting states, but the same noise prevents it from converging as precisely as aggregated GSP in later iterations.

To address the second issue, we recomputed the samples generated by the three methods using instead an incremental aggregation process, generating a summary sample for each iteration and target word by averaging all previous samples from all chains for that word, with iterations 1–6 treated as burn-in samples and hence discarded. The resulting samples are displayed in Figure 13. The aggregation process clearly improves sample quality for MCMCP and GSP (non-aggregated), but it does not fully solve MCMCP’s problem with poor sample quality. Though we do not have participant rating data for these aggregated samples, it is clear that MCMCP failed to converge on appropriate colors for chocolate, eggshell, and lavender. One might further criticize the lavender samples for being too red, the strawberry and sunset samples for not being red enough, and the lemon samples for being not yellow enough. It seems apparent that aggregating over trials does not necessarily resolve the performance issues of MCMCP in this paradigm.

To address the third issue, we tabulated samples from the previously described control experiment with different MCMCP proposal widths (Figure 14, Experiment 1d). There appears to be little difference in sample quality for the different proposal widths, but the extreme proposal widths do display certain characteristic behavior. As expected, we see that the MCMCP chains with the smallest proposal width (10) only make local adjustments to the color, meaning that once the chain gets close to an appropriate color category, it can be carefully tweaked to resemble this category as well as possible. However, these narrow-proposal chains often fail to approach the appropriate color category in the first place, even after 30 iterations. In contrast, the wide-proposal chains explore the color space quickly, but are unable to make subtle adjustments to match specific categories. The moderate proposal width of 30 provides some compromise between these two behaviors, and seems to be a sensible choice for the MCMCP-GSP comparison.
Figure 12: Mean sample ratings as a function of the participant time invested in chain construction (Experiments 1a–c, 1e), with time plotted on a log scale (95% confidence intervals over participants).

Figure 13: Colors derived by averaging raw samples from iteration seven onwards for the different sampling methods (Experiments 1a–c).
Appendix D Emotional prosody

D.1 Stimuli

The stimuli were created on the basis of three sentences from the Harvard sentences \cite{Harvard} recorded by a female speaker \cite{Speaker}. These sentences are phonologically balanced and semantically neutral. The stimulus space was then defined through seven continuous acoustic manipulations performed to these sentences. The manipulation were performed using the software Praat \cite{Praat} and the Python package Parselmouth \cite{Parselmouth}. Pitch (F0 contour) was extracted using a pitch floor of 100 Hz and ceiling of 500 Hz (default window size) using the command To Pitch in Parselmouth. Before proceeding we confirmed that all contours were free of any octave jumps. From the Sound and the Pitch object, we created a Manipulation object using the command To Manipulation. From the Pitch object we extracted the glottal pulses using To PointProcess. The manipulations were then performed in the following order:

1. *Pitch level*, shifting the pitch contour by a value in the range \([-37, 37]\) Hz.
2. *Pitch range*, scaling the original pitch range (expressed in Hz) by a value in the range \([20, 180]\)%, using the middle of the original pitch range as the center of the scaling operation.
3. *Pitch slope*, altering the original sentence’s pitch slope by a value in the range \([-37, 37]\) Hz. In our case, the reference sentences always began with a falling slope, and our manipulation was never severe enough to change them to a rising slope. Instead, a positive value of our pitch slope feature indicates a flattened contour, and a negative value indicates a steeply falling contour.

We manipulated pitch slope in the following way. We extracted the time of the first \((t_0)\) and last \((t_1)\) pitch values (ignoring unvoiced segments), and then edited the pitch contour by adding the following linear function \(f(t)\) to each pitch:

\[
f(t) = x \cdot \frac{t - t_0}{t_1 - t_0}
\]  

where \(t\) denotes the time of the point being edited and \(x\) denotes the feature value, ranging between \(-37\) Hz and \(37\) Hz.

We achieved this by creating an empty PitchTier object and populating it with the new contour using the command Add point. Finally we replace the old PitchTier in the Manipulation object with the new one using Replace pitch tier.
4. **F0 perturbation** is commonly measured as local frequency variation in the F0 contour (jitter), and corresponds approximately to the perceptual impression of hoarseness [12]. We modified F0 perturbation by converting the PointProcess object (representing the glottal pulses) to a Praat Matrix object (representing the time points of the pulses) using To Matrix. We changed the position of the pulses by applying the Praat formula `self + randomGauss(0, r)` where `r` was a number between 0 and 0.0001 determining the strength of the perturbation. The Matrix was converted back to a PointProcess with To PointProcess, and the glottal pulses in the Manipulation replaced using Replace pulses. This follows the algorithm proposed in [13].

5. **Duration**, allowed to change linearly from 80% to 120% from the original duration. To manipulate the duration we created an empty DurationTier object using the command Create DurationTier. At time 0 we placed a point with the value of the duration scalar using the command Add point 0 scalar. We then ran Replace duration tier to apply the changes. Note that changing the duration did not affect the overall pitch.

6. **Intensity variation**, corresponding to a periodic amplitude modulation of the signal. This manipulation was characterized by two parameters which constituted two independent dimensions of the stimulus space: amplitude modulation frequency (ranging from 0–5 Hz) and amplitude modulation depth (ranging from 0.01–10 dB). We implemented this using the operation ‘Vibrato and tremolo’ as defined in [13] and implemented in Parselmouth.

**D.2 Procedure**

The main chain-construction experiment (Experiment 2a) assigned each participant to one of three different emotions: happiness, sadness and anger. To ensure that all participants were familiar with the emotional concept we presented contexts that has been used in previous studies on emotional prosody [14, 15]:

**Anger:** Please think of a situation where you experienced a demeaning offense against you and yours. For example, somebody behaves rudely toward you and hinders you from achieving a valued goal. The situation is unexpected and unpleasant, but you have the power to retaliate.

**Happiness:** Please think of a situation where you made reasonable progress toward the realization of a goal. For example, you have succeeded in achieving a valued goal. Your success may be due to your own actions, or somebody else’s, but the situation is pleasant and you feel active and in control.

**Sadness:** Please think of a situation where you experienced an irrevocable loss. For example, you lose someone or something very valuable to you, and you have no way of getting back that what you want.

Each participant was randomly assigned to a different emotion (happy, sad, angry). After the headphone-screening task and a short demographic questionnaire, they took a practice trial to familiarize themselves with the slider. They then completed two within-participant chains corresponding to three alterations of each of the seven dimensions, alternating between both chains until both were complete. Each chain was initialized with the feature values of the reference sentence. In each trial the participant could chose from 25 stimuli synthesized from 25 equidistant points on the slider.

In the validation experiment (Experiment 2c), each participant rated stimuli for the three emotion words (happiness, sadness, anger) in three corresponding randomly ordered blocks. Each block contained 49 stimuli, which came in four types: a) raw samples from the GSP chains, b) samples derived by averaging the last three iterations of the GSP chains, c) the initial unchanged sentences, d) samples corresponding to random feature values. Participants were presented with the same emotional contexts as the participants in the chain-construction experiment, and responded using the same four-point scale as the other experiments (‘1. Not at all’, ‘2. A little’, ‘3. Quite a lot’, ‘4. Very much’).

We also conducted a control experiment where we switched from within-participant chains to across-participant chains, reducing the number of participants by approximately half because the original experiment proved to have more than sufficient power, and leaving all other experiment parameters
unchanged (Experiment 2b). Note that reducing the number of participants should not bias the validation ratings, which only used raw samples rather than samples created by aggregating over participants.

In the subsequent validation component (Experiment 2d), participants rated three blocks of 44 stimuli: 20 samples from the original within-participant chains, 20 samples from the new across-participant chains, 3 random samples, and one initial unchanged sentence. In all other regards this second validation was identical to the first validation.

D.3 Supplementary results

Figure 15A shows the results of the within- and across-participant comparison (Experiments 2a, 2b). The resulting feature values are broadly similar between these two experiments, suggesting that memory effects did not substantively contaminate our within-participant chains. This conclusion is supported by the validation experiment, which shows similar contrast scores for both within- and across-participant chains (Figure 15B, Experiment 2d).

Figure 15: Across-participant chains yield similar results to within-participant chains (Experiments 2a, 2b, 2d). A: Average parameter settings for across- and within-participant chains in iteration 20 (95% confidence intervals over chains). B: Mean validation contrast for different iterations (95% confidence intervals over participants). Contrast is defined as the difference between the rating for the target emotion and the mean rating for the non-target emotions. C: Average parameter settings for all iterations in within chains.

Figure 15C shows how mean feature values develop over the course of the within-participant experiment (Experiment 2a). Here we can see how most of the development of the feature values occurs over the first sweep of the feature vector (iterations 1–7), after which point the feature values stay broadly similar. Three audio examples from the same sentences in the final iteration of this experiment are included in the Supplementary Materials in the folder sound-examples-prosody: sad/happy/angry_final_sentence.wav and for reference the initial stimulus original_sentence.wav.

The Supplementary Materials can be accessed at https://www.aesthetics.mpg.de/forschung/forschungsgruppe-computational-auditory-perception/gsp.
GSP also allows us to investigate higher-order structure in perceptual representations. As an illustrative analysis, Figure 16 plots pairwise correlations for different features in the generated samples (Experiment 2a). For example, we see that duration and F0 perturbation were significantly correlated for sadness ($r = .28$) but not for the other emotions (anger: $r = -.03$, happiness: $r = .00$); in contrast we see that pitch level and pitch slope were positively correlated for all three emotions. These kinds of higher-order analyses provide a more expressive perspective on prosody features than previous research, which mainly focuses on the independent contributions of single features rather than interactions between features. We intend to explore these kinds of interactions more in future research.

![Figure 16: Pearson correlations between parameters in all three emotions (Experiment 2a).](image)

Appendix E   Musical chords

E.1 Supplementary methods

This study applied GSP to the perceived pleasantness of musical chords. Each of these chords comprises three tones, and is hence termed a triad. We represented each triad as a pair of numbers, following the ‘pitch chord type’ representation of [16], which represents each chord tone as a pitch interval in semitones from the bass (i.e., lowest) tone. This representation captures the sense in which human pitch perception is relative (i.e., pitches are heard relative to their recent auditory context) and logarithmic (i.e., perceived pitch distance is approximately proportional to the difference in the logarithm of the frequencies) [17]. Integer values in this representation correspond to the standard 12-tone equal temperament tuning system of Western music.

We generated chords using Tone.js, a Javascript library for synthesizing sounds in the client’s browser[9]. Each triad was synthesized as three simultaneous complex tones comprising 10 harmonics with amplitudes scaled by 12 dB/octave. These complex tones were presented with an ADSR envelope comprising a linear attack portion of 200 ms and a maximum amplitude of 1.0, an exponential decay portion lasting 100 ms taking the amplitude to 0.8, and a final exponential decay release portion lasting 1 s. The pitch of the bass tone was sampled uniformly and continuously in the logarithmic range G3–F4 (i.e., 196–349 Hz). The other two tones were specified by two continuous intervals in the range [0.5, 11], with the limits chosen such that the unison (0) and octave (12) were excluded, to

[9] https://tonejs.github.io/
Figure 17: Kernel density estimates of the four sets of KDE modes considered in the validation experiment for musical triads (Experiment 3b). The top five modes are indicated in red. The density values are computed relative to a uniform distribution.

We did however allow the two non-bass tones to overlap.

In each trial of the main experiment (Experiment 3a), participants were presented with the following prompt: ‘Adjust the slider to match the following word as well as possible: pleasant’. Releasing the slider prompted a chord to be played whose pitch intervals reflected the current position of the slider.

Two tones are said to share the same pitch class if they are separated by an integer multiple of 12 semitones (an octave).
Before beginning the main experiment, participants completed three training examples to familiarize themselves with the procedure.

Through this experiment we constructed 50 across-participant chains comprising 40 iterations. The starting seed for each chain was created by randomly sampling both intervals from a uniform distribution in the range \([0.5, 11]\). Each participant contributed up to 20 trials to these chains.

The validation experiment (Experiment 3b) collected ratings for various raw samples and kernel density estimate (KDE) modes, extracted from various iterations of the 50 chains constructed in the GSP experiment. The raw samples were represented by 16 experimental conditions corresponding to the iterations 0, 1, ..., 9 and 14, 19, ..., 39, with each condition being represented by 50 stimuli (1 stimulus from each chain). In addition to the raw sample conditions, there were four KDE conditions, with each KDE condition being represented by 5 modes, and the modes being extracted by aggregating data from iterations 0, 10 to 19, 10 to 29 and 10 to 39; here iteration 0 corresponds to the random seed (Figure 17). We used Gaussian kernels of width 0.175, and extracted modes using the kernel-based clustering algorithm of [18] as implemented in the ADPclust R package [19] with the number of clusters set \(\textit{a priori}\) to 20, and keeping the five resulting centroids with the highest kernel density. In total, this resulted in 20 experimental conditions comprising 820 stimuli in total.

In each trial of the validation experiment, the participant was assigned to a randomly chosen stimulus from one of the stimulus sets, and was asked to rate how pleasant that sound was on a four-level scale: ‘Not at all’, ‘A little’, ‘Quite a lot’ and ‘Very much’. Overall we collected 662 ratings for each experimental condition, with each participant contributing up to 80 ratings.

We should note that while in Gibbs sampling it is customary to consider samples in jumps of full coordinate sweeps, here we decided to aggregate data continuously, given the inherent symmetry between the two intervals, and especially as a means of mode estimation. We further exploited that symmetry by folding the data along the \(x = y\) line, since reordering of a pair of intervals does not alter the generated chord (Figure 18 shows the raw data where the \(x = y\) is clearly apparent, and Figure 17 show the folded kernels after re-ordering of the two intervals).

### E.2 Supplementary results

In the main paper, we mostly discussed the structure and validation of features and raw samples aggregated across various chains and iterations. Figure 18 complements this perspective by presenting...
Table of contents

1 Introduction
2 Aesthetic Perception
3 Methodology
4 Results
5 Conclusion

Appendices

Appendix F Faces

Figure 19: Marginal distributions for the combined intervals, using a sliding window of length 20 (Experiment 3a).

The trajectory of a typical chain (Experiment 3a). It is clear that the dynamics are far from an optimization regime, where one would expect to see small and converging updates toward some local optimum [e.g. 20]. Instead, the trajectories illustrate the sampling regime of GSP, characterized by big leaps and lack of convergence, scanning the various regions of the space.

Figure 19 shows the behavior of the marginal distribution for both intervals combined, computed over a sliding window of length 20 with Gaussian kernels (bandwidth = 0.175 semitones, Experiment 3a). As can be seen, two strong modes emerge at the perfect fifth (7) and the major third (4) (and other integer peaks), along with dips at the semitone (1) and tritone (6), reflecting the standard Western tonal hierarchy.

Audio samples of the top 15 KDE modes extracted from iterations 0 (random) and 10–39 can be found in the Supplementary Materials in the folder sound-examples-musical-triads, with the modes arranged in descending order of density (random_seed_top_modes_iter_0_0.wav and top_modes_iter_10_39.wav).

Appendix F Faces

F.1 Supplementary methods

This study used the ‘StyleGAN’ model of [21, 22] pretrained on the FFHQ dataset of faces from Flickr [21]. This model is a generative adversarial network, comprising a latent vector $z$ sampled from a probability distribution $p(z)$, an input layer that takes a constant input $y_0$, and the other layers $y_i$ taking the previous layer and a non-linear function of $z$ as an input:

$$y_i = G_i(y_{i-1}, w), \quad w = M(z),$$

where $M$ is an 8-layer multilayer perceptron and the output layer $y_L$ corresponds to an RGB image. The study depended on participants interactively manipulating principal components of the $w$ vector using a slider. We achieved this by creating an API that took as input a random seed for the latent vector $z$, a vector of principal component values for $w$, and the index of the principal component to be manipulated by the slider. The API then returned a video where the active principal component
was incrementally modified through a specified number of standard deviations about the mean, with this API building on code released in [23]. The resulting video was then streamed to the participant’s local computer, with the slider selecting between different frames of the video. We hosted the API on an AWS EC2 instance fitted with an NVIDIA K80 GPU.

An important technical issue concerned ensuring that participants didn’t have to wait for the relatively slow stimulus generation process. We therefore generated stimuli asynchronously in advance of a given experimental trial, with participants being randomly assigned to the pool of currently available stimuli for each trial. Aggregating multiple responses per step of the GSP process helped in this regard, meaning that a higher throughput of participants could be sustained for a given rate of stimulus production.

The main experiment (Experiment 4a) evaluated six adjectives which we thought could elicit meaningful perceptual associations: ‘attractive’, ‘fun’, ‘intelligent’, ‘serious’, ‘trustworthy’, and ‘youthful’. Three across-participant chains were constructed for each of these adjectives, each of length 50 plus the initial random state, resulting in a total of 18 chains. Each step in the chain received five responses from five different participants, which were then aggregated using the arithmetic mean.

Participants were recruited from Amazon Mechanical Turk as before with the stipulation that they be resident in the US. All participants were pre-screened with the color vocabulary task used previously for the color experiment. After completing a short demographic questionnaire, they took six practice trials to familiarize themselves with the task, then proceeded to the main experiment, where they completed up to 18 trials (one from each chain).

The validation experiment (Experiment 4d) recruited participants in the same manner, and had the participants rate all generated samples from iterations 1–10, 20, 30, 40, and 50. Each participant contributed 80 ratings, under the constraint that they never rated the same sample twice, and with participants being assigned to stimuli such that the number of ratings accumulated equally across stimuli. Data collection was continued until all samples had been rated at least 50 times.

We additionally conducted two control experiments exploring a) the effect of aggregating with KDE modes instead of arithmetic means and b) the robustness of the paradigm to including non-US participants. The first control experiment (Experiment 4b) implemented KDE using a Gaussian kernel with a standard deviation of 0.5 in units of PCA standard deviations. We initialized each chain using the same starting values as the original experiment to maximize comparability between the two experiments. The second control experiment (Experiment 4c) replicated the first control experiment, but removed the recruitment criterion originally stipulating that all participants must be US residents. The resulting participant group was dominated by Indian (c. 50%) participants but also included a high proportion of US participants (c. 40%). Unlike the previous control experiment, all chains were initialized with fresh random values.

We followed up these two control experiments with an updated validation experiment (Experiment 4e) evaluating samples generated from the original experiment (US-only, mean aggregation) as well as the two control experiments. This validation experiment had the same design as the original validation experiment, but including raw samples from all three experiments, and decreasing the target number of ratings per sample to 25. The participant group for this experiment was recruited using the US-only criterion.

F.2 Supplementary results

Example raw samples and validation results for the main experiment (Experiment 4a, 4d) are shown in Figure 4 of the main paper. Figure 20 illustrates the raw samples in more detail, displaying iterations 0–10, 20, 30, 40, 50 from one chain for each target word. It is clear from both the validation results and the raw samples that the chains make good progress towards the target category already by the end of the first sweep (10 iterations), and sometimes the resemblance to the target category does not improve noticeably after this point. However, this does not mean that the process converges to a static image after this point: instead, there is a moderate amount of variety in the subsequent faces (see also Figures 22 and 23). The process is therefore still somewhat in the stochastic sampling regime rather than the deterministic optimization regime.

Validation results for the two control experiments (Experiments 4b, 4c) and the main experiment (Experiment 4a) are shown in Figure 21 (Experiment 4e). The broad trends in the ratings are replicated
Figure 20: Raw samples from six GSP chains in Experiment 4a (US-only participants, mean aggregation).

Figure 21: Validation results for Experiments 4a, 4b, 4c, as produced in Experiment 4e (95% confidence intervals over participants).

across the three experiments: typically almost all of the improvement comes in iterations 1–10, with ratings staying broadly stable after this point. All three experiments struggle to capture trustworthiness, which is clearly a particularly subjective judgment to make. Interestingly, there is no evidence that KDE peak-picking outperforms the arithmetic mean as an aggregation technique. Inspecting the raw data and the density estimates, this did not seem to be a consequence of poorly chosen kernel width or artifacts in the density estimation process. Instead, it seems that the participants’ conditional distributions could typically be approximated well by a unimodal distribution, and hence averaging was a sensible aggregation method.
Raw samples of six chains from each control experiment are displayed in Figures 22 and 23. It is important not to read too much into these raw samples, as they ultimately come from stochastic distributions and will vary over repeated runs. However, we did notice some suggestive differences between the final samples of the non-US chains and those of the US chains. Most salient was the fact that all US chains for ‘intelligent’ finished with a Caucasian man, whereas the three final states of the non-US chains included both a woman and a non-Caucasian man. We also noticed that the non-US chains were the only ones to include a man as the final ‘attractive’ sample. While some of this variation will be due to chance, the remaining variation will presumably reflect different stereotypes held by the different participant groups. It would be very interesting to explore these different stereotypes in a more systematic way.

Figure 22: Raw samples from six GSP chains in Experiment 4b (US-only participants, KDE mode aggregation).

Appendix references

[1] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, “Motion illusions as optimal percepts,” *Nature Neuroscience*, vol. 5, no. 6, pp. 598–604, 2002.
[2] X.-X. Wei and A. A. Stocker, “A Bayesian observer model constrained by efficient coding can explain ‘anti-Bayesian’ percepts,” *Nature Neuroscience*, vol. 18, pp. 1509–1517, 2015.
[3] A. N. Sanborn, T. L. Griffiths, and D. J. Navarro, “A more rational model of categorization,” in *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (R. Sun and N. Miyake, eds.), pp. 726–731, Cognitive Science Society, 2006.
[4] D. McFadden, “Conditional logit analysis of qualitative choice behaviour,” in *Frontiers in Econometrics* (P. Zarembka, ed.), pp. 105–142, New York, NY: Academic Press, 1974.
[5] K. E. Train, *Discrete choice methods with simulation*. Cambridge University Press, 2009.
[6] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, “Headphone screening to facilitate web-based auditory experiments,” *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.
[7] J. H. Clark, “The Ishihara Test for color blindness,” *American Journal of Physiological Optics*, vol. 5, pp. 269–276, 1924.
[8] “IEEE recommended practice for speech quality measurements,” tech. rep., IEEE, 1969. ISBN: 9781504402743.
Figure 23: Raw samples from six GSP chains in Experiment 4c (global participant group, KDE mode aggregation).

[9] P. Demonte, “HARVARD corpus speech shaped noise and speech-modulated noise for SIN test,” 2019. Publisher: University of Salford.

[10] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program],” Version 6.0.37, retrieved 3 February 2018, http://www.praat.org/, 2018.

[11] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” Journal of Phonetics, vol. 71, pp. 1–15, 2018.

[12] I. R. Titze, Y. Horii, and R. C. Scherer, “Some technical considerations in voice perturbation measurements,” Journal of Speech, Language, and Hearing Research, vol. 30, pp. 252–260, June 1987.

[13] R. Corretge, “Praat vocal toolkit.” http://www.praatvocaltoolkit.com, 2020.

[14] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner, “The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures,” Nature Human Behaviour, vol. 3, pp. 369–382, Apr. 2019.

[15] P. Laukka, H. A. Elfenbein, N. S. Thingjum, T. Rockstuhl, F. K. Iraki, W. Chui, and J. Althoff, “The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features,” Journal of Personality and Social Psychology, vol. 111, Nov. 2016.

[16] P. M. C. Harrison and M. T. Pearce, “Representing harmony in computational music cognition,” PsyArXiv, 2020.

[17] T. Stainsby and I. Cross, “The perception of pitch,” in The Oxford handbook of music psychology (S. Hallam, I. Cross, and M. Thaut, eds.), pp. 47–58, New York, NY: Oxford University Press, 2009.

[18] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” Science, vol. 344, no. 6191, pp. 1492–1496, 2014.

[19] X.-F. Wang and Y. Xu, “Fast clustering using adaptive density peak detection,” Statistical methods in medical research, vol. 26, no. 6, pp. 2800–2811, 2017.

[20] N. Jacoby and J. H. McDermott, “Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction,” Current Biology, vol. 27, no. 3, pp. 359–370, 2017.
[21] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410, 2019.

[22] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” arXiv, 2019.

[23] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, “GANSpace: Discovering interpretable GAN controls,” arXiv, 2020.