Cognition 226 (2022) 105196

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Range-frequency effects can explain and eliminate prevalence-induced concept change

David E. Levari [a,b,*]

[a] Harvard Business School, Boston, MA, United States of America
[b] Department of Psychology, Harvard University, Cambridge, MA, United States of America

ARTICLE INFO

Keywords:
Prevalence
Context
Reference dependence
Social cognition
Perception

ABSTRACT

Why would concepts seem to grow when their instances become rare? Human observers can respond to decreases in stimulus prevalence by expanding their conceptual boundaries of those stimuli. This *prevalence-induced concept change* may have serious social consequences, since many real-world detection tasks demand consistent judgments over time. The current work aims to identify the computational process that produces prevalence-induced concept change. I review some plausible models from the cognitive and social sciences that could account for this phenomenon, and then use trial-level computational modeling to see how well each model predicts actual human data, finding that they are best explained as a range-frequency compromise in judgment. Finally, I test an intervention that successfully eliminates prevalence-induced concept change by making stimuli more intense as they become rare.

1. Introduction

How do subjective judgments of stimuli change as they become more or less common? Levari et al. (2018) documented a peculiar trend in judgments of colors, threatening faces, and moral violations – as stimuli in those categories decreased in prevalence, human observers expanded their boundaries to include a wider range of stimuli. For example, when the prevalence of blue dots (relative to non-blue dots) was reduced from 50% to 6%, participants began identifying colors as blue that they had previously judged to be non-blue. Conversely, when the prevalence of blue dots was increased, participants narrowed the range of colors they identified as blue. Adjusting perceptions and judgments based on available resources is an adaptive behavior seen in many domains and animal species (e.g. Hayden, 2018; McNair, 1982; Wolfe, 2013). However, the phenomenon Levari and coauthors examined, which they called *prevalence-induced concept change*, has troubling implications for domains in which consistency is important. As the authors suggested, "…[when] yellow bananas become less prevalent, a shopper's concept of 'ripe' should expand to include speckled ones, but when violent crimes become less prevalent, a police officer's concept of 'assault' should not expand to include jaywalking."

If prevalence-induced concept change is sometimes undesirable, can it be prevented? Levari et al. (2018) attempted several experimental interventions to reduce or eliminate the phenomenon in judgments of color. In one study, they warned participants in advance about the prevalence change in the study. In another, they explicitly asked participants to stay consistent over time in their judgments, and not to change which colors they called blue. Another study offered financial incentives for participants who could stay consistent over time. Finally, the experimenters tried decreasing the prevalence of blue dots abruptly instead of gradually, to make it more noticeable. None of these interventions were successful, perhaps in part because the mechanism driving the phenomenon was unknown. In this paper, I use trial-level computational modeling (Daw, 2011) to search for a mechanism that predicts and explains prevalence-induced concept change. I then use the results of that model to design a new intervention to reduce or eliminate the effect.

1.1. Overview

What computational mechanism could produce prevalence-induced concept change? Levari and coauthors (2018) speculated that the phenomenon was driven by contextual comparisons, in which observers compared the intensity of the current stimulus to recently seen stimuli. Such a comparison would lead participants to judge a dot as more blue when it was preceded by very non-blue dots (low prevalence) than when

* Corresponding author at: Harvard Business School, Boston, MA, United States of America.
E-mail address: dlevari@fas.harvard.edu.

https://doi.org/10.1016/j.cognition.2022.105196
Received 27 August 2021; Received in revised form 27 May 2022; Accepted 31 May 2022
Available online 6 June 2022

preceded by very blue dots (high prevalence). Psychologists, neuroscientists, and economists have extensively studied these kinds of contextual influences on human judgment and perception (Bhui, Lai, & Gershman, 2021; Schwartz, Hsu, & Dayan, 2007; Spektor, Bhatia, & Gluth, 2021; Summerfield & de Lange, 2014). In this section, I will review a few potential mechanisms drawn from prior research on contextual effects in judgment and perception that predict behavioral effects similar to prevalence-induced concept change.

1.1.1. Bayesian models of perception and judgment

Cognitive scientists have done extensive work exploring which mental operations in the brain can be better understood with Bayesian models of cognition (Griffiths, Kemp, & Tenenbaum, 2008). In such models, the brain uses prior knowledge from past events to inform perceptions, predictions, and judgments. Some classic exercises in Bayesian inference use prevalence to illustrate how prior knowledge can usefully improve predictions. In one example, the "cab problem" (Birnbaum, 1983), accurately weighting eyewitness testimony about the color of a taxicab involved in a hit-and-run accident requires consideration of the base rates of different colors of cars.

In one simple version of Bayesian probability estimation, imagine that an observer is trying to guess the probability that a given dot is blue. The prior used to update the posterior probability that the dot is blue is the current base rate of blue dots. When blue dots are rare, the prior goes down, as does the posterior probability. As a result, observers should be biased against finding more blue dots – a reversal of the typical prevalence-induced concept change effect. However, some Bayesian models of perception and judgment are more compatible with Levari et al.'s findings. In a second kind of model, imagine an observer who attempts to infer the color of a dot by comparing it to the average color of dots in the environment, and who uses some form of Bayesian updating to adapt their estimation of that average to better match recently seen dots. Such models (e.g. Wei & Stocker, 2015) would predict, as Levari et al. (2018) find, that observers are more likely to call dots blue when bluer dots become rare. Similarly, models of Bayesian learning that use beta distributions to inform inferences about color boundaries based on binomial outcomes (e.g. J. Feldman, 2021) could also account for such effects, by biasing inferences towards previously common intermediate colors, rather than rare and extreme colors.

1.1.2. Sensory aftereffects and value adaptation

Repulsive aftereffects are a celebrated phenomenon in vision research, in which perceptions of stimuli are biased away from the average of recently seen stimuli. Prolonged exposure to photographs of artificially widened faces causes normal faces to temporarily seem too narrow (Rhodes, Jeffery, Watson, Clifford, & Nakayama, 2003). Similar effects have been documented for colors (Webster, 1996), motion (Anstis, Verstraten, & Mather, 1998; Mather, Pavan, Campana, & Casco, 2008), and spatial orientation (Paradiso, Shimojo, & Nakayama, 1989), though there are also extensive examples of the opposite phenomenon, sometimes called an attractive aftereffect or positive serial dependency (e.g. Cicchini, Mikellidou, & Burr, 2017; Fornaciai & Park, 2018; Manassi, Liberman, Kosovicheva, Zhang, & Whitney, 2018), as well as a growing understanding of what drives these opposing effects in seemingly similar contexts (e.g. Fritsche, Mostert, & de Lange, 2017). Prevalence effects in subjective judgment could be driven by the same mechanisms that produce repulsive aftereffects, a possibility raised by Vickers and Leary (1983) and foreshadowed by Adaptation-level Theory (Helson, 1964). One such candidate mechanism is normalization (Heeger, 1992), a computation in which input to a given neuron is divided by the pooled activity of similarly tuned neurons. Normalization has been used to successfully explain contextual effects in vision (Carandini, Heeger, & Movshon, 1997) as well as choice and valuation (Louie, Khaw, & Glimcher, 2013; Webb, Glimcher, & Louie, 2020), and is often considered an important example of efficient coding in neural architecture (Attneave, 1954; Barlow, 2001), because storing the differences between values eliminates redundancies compared to storing the values themselves.

1.1.3. Range-frequency theory

Prevalence-induced concept change is also in line with the predictions of Range-Frequency Theory (Parducci, 1963, 1965), which describes the influence that particular distributions of stimuli can have on judgments of those stimuli. As applied to categorical judgments (Parducci & Wedell, 1986), it posits that the boundaries between categories (e.g. "yes/no", "blue/purple", a 5 point Likert scale) are implicitly constructed by two contextual factors that influence the subjective value of stimuli. The first factor is *range*, or the minimum and maximum stimulus values present. The second factor is *frequency*, or the distribution of other stimulus values present in the observer's current context. Subjective judgments are influenced by a weighted compromise between these two factors.

Range-Frequency Theory (RFT) is often used to model the behavior of observers either at a low frequency or a high frequency, but less often of observers transitioning between these states. The prevalence manipulation employed in Levari et al. (2018) directly manipulated what RFT would call the frequency parameter, or the prevalence of stimuli. By contrast, the range of possible stimuli was fixed in each experimental session. In such a scenario, RFT predicts that as the frequency of stimuli decreases, the subjective intensity of a given stimulus should increase, because its rank intensity (e.g. whether it is the 5th or the 50th most intense stimulus seen recently) is now greater than it was when the frequency was high. In the case of deciding whether a dot is blue or purple, as blue dots become more rare, a given blue dot should seem relatively more blue than it did previously. This is because that dot is now more blue than other recently seen dots, compared to when blue dots were common.

1.1.4. Adjudicating between different computational models

In trial-level computational modeling (Daw, 2011; Wilson & Collins, 2019), each model under consideration is expressed with an algorithm that can generate a response on each trial of a task. These responses and the likelihood function that generated them are then compared to actual human responses. The goal is to determine which model among those being compared most accurately describes real human behavior on each individual trial of the task, rather than in aggregate behavior of groups of participants. This approach is particularly useful when comparing similar models, as is the case here, since the models described above can all predict prevalence effects that broadly match the direction of those found in Levari et al. (2018). The potential similarity of these models' predictions also makes it important to determine whether they can be reliably distinguished using a computational model comparison. This danger can be reduced with model-recovery simulations (e.g. Edmunds, Milton, & Wills, 2018), in which each model being compared is used to simulate responses, which are then fit by each model in turn. Ideally, each model is able to fit "recover" its own simulated responses better than the other models, making misleading conclusions based on computational modeling methods less likely.

2. Study 1: Replication of Levari et al. (2018) study 1

2.1. Overview

In a direct replication of Levari et al. (2018) Study 1, I recruited an online sample of participants to view a series of dots on a computer screen and identify each dot as either blue or not blue. After many trials, the prevalence of blue dots decreased for some participants. I then used trial-level computational modeling to see which of several candidate mechanisms best predicted the actual human data in the study. This and all subsequent studies were preregistered on AsPredicted.org (see Supplementary Materials for link), and approved by the Harvard University Committee on the Use of Human Subjects. Informed consent was

provided by all participants.

2.2. Methods

2.2.1. Participants

Participants were 47 users of the online survey platform Amazon Mechanical Turk, recruited via CloudResearch.com (Litman, Robinson, & Abberbock, 2017) (24 males, 22 females, 1 prefer not to answer, M_{age} = 42.28 years, SD = 13.61 years) who were paid $3 USD for their participation. Minimum required sample size in this and all studies was determined using the R package *simr* (Green & MacLeod, 2016) to reach 90% statistical power to detect effects comparable to those reported in Levari et al. (2018).

2.2.2. Procedure

Participants were told that a series of colored dots would appear on the screen, one at a time, and that their task was to decide whether each dot was blue or not blue, and to indicate their decision by pressing one of two keys on the keyboard ("f" for not blue, "j" for blue).

On each trial, a colored dot appeared on a solid gray background. The color of the dot varied across trials from very purple (61% blue, RGB 99–0-155) to very blue (100% blue, RGB 0–0-254). Each dot appeared on the screen for 500 milliseconds and was then replaced by a question mark, which remained on the screen until participants pressed one of the response keys. Participants were told that there would be 800 trials divided into 16 blocks, and that the prevalence of blue dots might vary across blocks. Specifically, they were told that some blocks "may have a lot of blue dots, and others may have only a few." Participants completed 10 practice trials to ensure they understood the procedure, and then completed 800 test trials. To help participants remain attentive, I allowed them to take a brief break every 50 trials. Stimulus presentation was programmed using *jsPsych* (de Leeuw, 2015).

I created two conditions by dividing the color spectrum into two halves that I will refer to as the "purple spectrum" (RGB 99–0-155 through RGB 51–0-204) and the "blue spectrum" (RGB 50–0-205 through RGB 0–0-254). Half the participants were randomly assigned to the *stable* condition. In this condition, I determined the color of the dot shown on each trial by randomly sampling the two spectra with equal probability from a uniform distribution. I will refer to the probability that a dot was sampled from the blue spectrum as the *signal prevalence*. In the stable condition, the signal prevalence on trials 1–800 was 50%. The remaining participants were assigned to the *decreasing* condition. In this condition, I sampled the two spectra with unequal probability on some trials. Specifically, in the decreasing condition the signal prevalence was 50% on trials 1–200; 40% on trials 201–250; 28% on trials 251–300; 16% on trials 301–350; and 6% on trials 351–800. After completing the identification task, participants completed a questionnaire asking some basic demographics and their impressions of the task. The complete text of the task instructions and all questions is available in Supplemental Appendix A.

2.3. Results

Following my preregistration, I did not exclude any participants from my analysis. To find out whether the decrease in the prevalence of blue dots cause participants to call a wider range of colors blue, I fit a binomial generalized linear mixed model to my data in R (R Core Team, 2020) using the package *lme4* (Bates, Mächler, Bolker, & Walker, 2015). The dependent variable was the participant's *identification* of a dot as blue or not blue. The independent between-participants variable was the participant's *condition* (stable or decreasing). The independent within-participants variables were (a) the dot's RGB value or what I will call its *actual color* (which ranged from 60% blue to 100% blue, coded as 1-100) and (b) the *trial number* (which ranged from 1 to 800). I included condition, trial number, and actual color (and all interactions between them) as fixed effects in my model. I included a random intercept term for participants (who may have entered the study with different thresholds) and allowed slopes to vary randomly across trial number for each participant (since responses over time may change in different directions or by different amounts from participant to participant). The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 594.43$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 292.37$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and actual color significantly improved model fit, $\chi^2(1) = 191.10$, $p < 0.001$.

The generalized linear mixed model revealed that a Condition X Actual Color X Trial Number interaction predicted participants' identifications, $b = 14.80$, $SE = 0.88$, $z = 16.75$, $p < 0.001$, *95% CI*$_b$ [13.07, 16.53], $R^2_{GLMM(c)} = 0.86$. Fig. 1 shows the percentage of dots at each point along the continuum that participants identified as blue on the initial 200 trials and on the final 200 trials. The two curves in the left panel are nearly perfectly superimposed, indicating that participants in the stable condition were just as likely to identify a dot as blue when it appeared on an initial trial as when it appeared on a final trial. But the two curves in the right panel are offset, indicating that participants in the decreasing condition were more likely to identify dots as blue when those dots appeared on a final trial than when those dots appeared on an initial trial. In other words, when the prevalence of blue dots decreased, participants called a wider range of colors blue.

3. Computational modeling of prevalence-induced concept change

Here I report a computational modeling approach to prevalence-induced concept change in color judgment, in which several possible cognitive mechanisms for the phenomenon are compared to see which model best predicts the behavioral results from Study 1. Aside from the first model, which serves as a control, each model I test employs some way of using local prevalence to update either the observer's categorical boundary between colors, or the subjective evaluation of color intensities. I first use each model to simulate data in the color identification task in Study 1, and test whether each model can not only accurately estimate the true simulation starting parameters, but also recover its own simulated responses better than the other models. Then, I fit each model to the data from Study 1, using Bayesian Model Selection to determine which model best accounts for actual human responses in the task.

3.1. Model specifications

3.1.1. Model 1: Normal CDF

This model serves as a control since it has no ability to adapt subjective intensity of stimuli or threshold values between categories. It implements a kind of classic psychophysical categorical perception (e.g. Decarlo, 2013; Feldman, Griffiths, & Morgan, 2009), in which the probability that the intensity x of the current stimulus exceeds the intensity τ is determined by a normal cumulative distribution function (CDF), where x is the intensity of the current stimulus and σ is the standard deviation:

$$P(x > \tau) = \Phi\left(\frac{x-\tau}{\sigma}\right) \tag{1.1}$$

The decision rule as to whether a given color x belongs to concept C (blue or not blue) is governed by the following probabilities:

$$P(C|x) = \begin{cases} \Phi\left(\frac{x-\tau}{\sigma}\right) \text{ if } C = blue \\ 1 - \Phi\left(\frac{x-\tau}{\sigma}\right) \text{ if } C = not\ blue \end{cases} \tag{1.2}$$

The free parameters in the model are τ (the decision threshold) and σ (the standard deviation of the normal CDF).

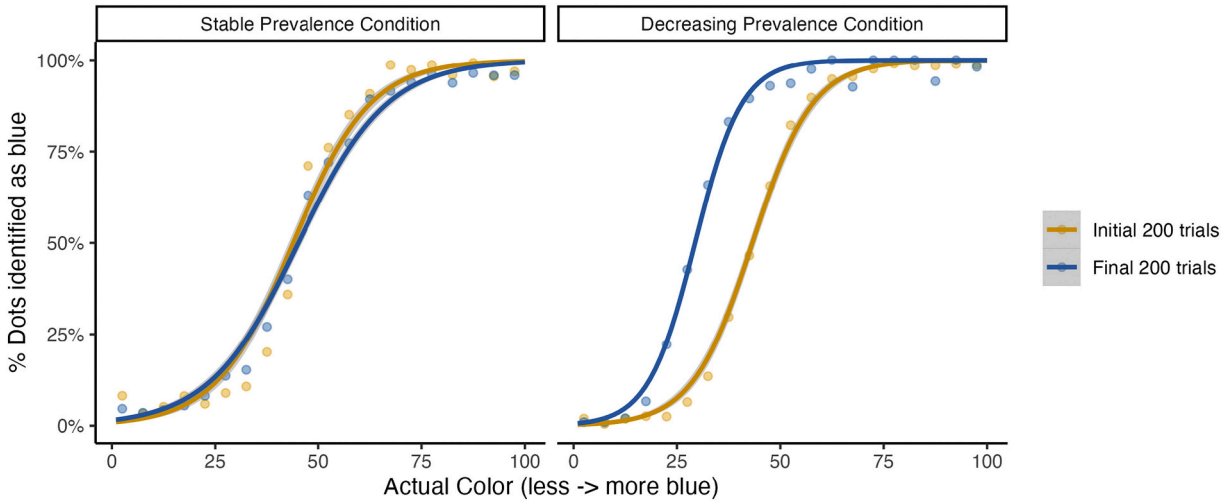


Fig. 1. Results of Study 1.
The x axis shows the dot's objective color, and the y axis shows the percentage of trials on which participants identified that dot as blue. Fitted lines were computed as binomial GLMs.

3.1.2. Model 2: Bayesian MAP Estimator

The second model builds on the basic model by implementing dynamic updating of the decision threshold through Bayesian max*imum* a posteriori (MAP) estimation (Griffiths & Yuille, 2008). In this model, the agent is attempting to infer the posterior probability around the threshold τ given the most recent observed trial x_i:

$$P(\tau|x_i) = P(x_i|\tau)P(\tau) \tag{2.1}$$

P(τ), the prior, is defined as a normal probability distribution function with starting mean of τ_0 and deviation of σ, or the initial starting threshold and variance values that the observer enters the task with. P($x_i|\tau$) is the likelihood of observing the current color given τ, in this case via a normal distribution with mean τ and standard deviation σ at value x_i. The entire posterior probability can be reformulated as follows:

$$P(\tau|x_i) = \mathcal{N}\left(x;\tau_i,\sigma^2\right)\mathcal{N}\left(\tau;\tau_{i-1},\sigma^2\right) \tag{2.2}$$

Once this posterior probability P($\tau|x_i$) is obtained, the maximum value τ is estimated from the posterior normal probability distribution function:

$$\tau = max\left(\mathcal{N}\left(x_i;\tau,\sigma^2\right)\right) \tag{2.3}$$

This value is then passed to the normal CDF function as implemented in Model 1 (Eqs. (1.1) and (1.2)), in order to classify the current color depending on whether it is greater or less than τ. The free parameters in the model are τ_0 (the initial decision threshold), and σ (the standard deviation of the normal CDF).

3.1.3. Model 3: Range-Frequency model

This model fixes τ and σ at their initial starting values, and attempts to determine the subjective intensity y_i of the current stimulus x_i within the local context *k* of recently observed values ($x_1, \ldots, x_n$). The subjective intensity is calculated with the range (minimum and maximum values of *x*) and frequency (the ordinal rank of the current stimulus within all values in *k*). The tradeoff between the influence of range and frequency on y_i is determined by the weighting parameter *w*.

$$y_{ik} = w\left[\frac{x_i - x_{min,k}}{x_{max,k} - x_{min,k}}\right] + (1-w)\left[\frac{rank_{ik} - 1}{n_k - 1}\right] \tag{3.1}$$

The subjective color value y_{ik} is then passed to the normal CDF function as implemented in the basic model (Eqs. (1.1) and (1.2)), where it serves as the intensity of the current stimulus (*x*). The free parameters in the model are n_k (the maximum number of trials included in the local context *k*), τ_0 (the decision threshold) and σ (the standard deviation of the normal CDF). In this implementation, the range-weighting parameter *w* is fixed at 0.5.[1]

3.1.4. Model 4: Moving Window

This model attempts to adapt the categorizations of a normal CDF to local context by updating the decision threshold τ using recently seen trials. Specifically, the current threshold τ_i is an incremental update of the previous threshold τ_{i-1} updated based on an exponentially-weighted moving average of past trials:

$$\tau_i = \tau_{i-1} + \alpha(x_{i-1} - \tau_{i-1}) \tag{4.1}$$

where τ_i is the current threshold, τ_{i-1} is the estimated threshold from the previous trial, α is a scaling parameter reflecting how much to update the current threshold based on past information (similar to a learning rate in reinforcement learning), and *x* is the observed color on a given trial. The current threshold value is then passed to the normal CDF function as implemented in Model 1 (Eqs. (1.1) and (1.2)), in order to classify the current color depending on whether it is greater or less than τ. The free parameters are α (the learning rate), τ_0 (the initial decision threshold), and σ (the standard deviation of the normal CDF).

3.1.5. Model 5: Adaptive value coding

This model implements a simplified value normalization algorithm (Khaw, Glimcher, & Louie, 2017). It updates the subjective value of the currently observed stimulus, rather than the decision threshold used to identify stimuli. The value of the stimulus *x* on the current trial *i* is divided by a summation of the values of the past *n* trials indexed by *k*, times a scaling parameter α. This value is then scaled by factor K, which represents gain.

[1] I fixed the range-weight at 0.5 to reduce model complexity. To check that this was a reasonable choice in model fitting, I also implemented a modified version of Model 3 that added range-weight as a free parameter that was separately estimated for each participant. Using this model, the median range-weight for participants in Study 1 was 0.47 (mean = 0.41, SD = 0.17). Further, Model 3 (with range-weight fixed at 0.5) strongly outperformed the modified model when fit to the data from Study 1. In other words, adding range-weight as a free parameter did not improve Model 3's ability to predict actual behavior. This is likely because range-weight did not vary strongly between subjects, and because of the increased model complexity from adding an additional parameter, which is penalized in Bayesian Model Selection.

$$y_i = K\frac{x_i}{1+\alpha\sum_{k=1}^{n}x_{i-k}} \tag{5.1}$$

The subjective color value y_i is then passed to the normal CDF function as implemented in Model 1 (Eqs. (1.1) and (1.2)), where it serves as the intensity of the current stimulus (x). The free parameters in the model are n (the maximum number of trials included in the local context), α (the contextual scaling parameter), τ (the decision threshold), and σ (the standard deviation of the normal CDF). Following Louie et al. (2013), the gain parameter K is fixed at 100 to generate plausible subjective intensity values that can be passed to the normal CDF function.

3.1.6. Model 6: Range-only model

This model implements the range component of Range-Frequency theory from Eq. (3.1), with no tradeoff parameter (since there is no frequency component present).

$$y_{ik} = w\left[\frac{x_i - x_{min,k}}{x_{max,k} - x_{min,k}}\right] \tag{6.1}$$

The subjective color value y_{ik} is then passed to the normal CDF function as implemented in Model 1 (Eqs. (1.1) and (1.2)), where it serves as the intensity of the current stimulus (x). The free parameters in the model are n_k (the maximum number of trials included in the local context k), τ (the decision threshold), and σ, (the standard deviation of the normal CDF).

3.1.7. Model 7: Frequency-only model

This model implements the frequency component of Range-Frequency theory from Eq. (3.1), with no range-weight parameter (since there is no range component present).

$$y_{ik} = \left[\frac{rank_{ik} - 1}{n_k - 1}\right] \tag{7.1}$$

The subjective color value y_{ik} is then passed to the normal CDF function as implemented in the Model 1 (Eqs. (1.1) and (1.2)), where it serves as the intensity of the current stimulus (x). The free parameters in the model are n_k (the maximum number of trials included in the local context k), τ (the decision threshold), and σ (the standard deviation of the normal CDF).

3.1.8. Model 8: Beta-Binomial model

This model implements dynamic updating of the decision threshold via Bayesian learning with a Beta-Binomial model (e.g. Feldman, 2021; Stankevicius, Huys, Kalra, & Seriès, 2014). In this model, as in Model 2, (Eq. (2.1)), the agent is attempting to infer the posterior probability around the threshold τ given the most recent observed trial x_i. P(τ), the prior, is defined as a beta distribution with parameters A and B:

$$P(\tau) = \frac{1}{B(A,B)}x^{A-1}(1-x)^{B-1} \tag{8.1}$$

This prior is conjugate to the binomial likelihood and has starting values of A_0 and B_0, representing initial evidentiary strength for blue and purple trials, respectively. The posterior probability $P(\tau|x_i)$ is also a beta distribution, in which either A or B is incremented by 1 on each trial, depending on whether the most recent dot was classified as blue or purple, respectively.

The maximum value τ is estimated from the posterior beta distribution, and then passed to a normal CDF function as implemented in Model 1 (Eqs. (1.1) and (1.2)), in order to classify the current color depending on whether it is greater or less than τ. The free parameters in the model are A_0 (the initial A value), and B_0 (the initial B value), and σ (the standard deviation of the normal CDF).

3.2. Generation of simulated data

Data for 20 simulated agents in the color identification task from Study 1 were generated in MATLAB from each of the eight models being tested. The true parameters for each agent were randomly sampled from uniform distributions. In Models 1–7, τ/τ_0 were sampled from the uniform distribution from 0.001 to 100, denoted as U(0.001,100), as was σ in Models 1–8. In the Moving Window and Adaptive Value Coding models, α was sampled from U(0.001,1). In the Moving Window, Range-Frequency, Adaptive Value Coding, Range Only, and Frequency Only models, n/n_k was sampled from U(1,90). In the Beta-Binomial model, A_0 and B_0 were each sampled from U(0.01,100).[2] Each agent completed 800 trials. I then used a trial-level model fitting procedure (Daw, 2011) with maximum likelihood estimation to recover the true generative model and parameters of the simulated data.

3.3. Recovery of simulated data

Optimized parameters for each simulated agent were estimated with the MATLAB package *mfit* (Gershman, 2015). Five starting values were uniformly sampled for each parameter, with the same bounds as used for data generation (described above).

To test the ability of each model to recover its own simulated responses, I calculated a protected exceedance probability for the simulated responses from each of the eight models, fit in turn by each of the eight models (a total of 64 model fits). The protected exceedance probability is a value from Bayesian Model Selection (Rigoux, Stephan, Friston, & Daunizeau, 2014) quantifying the likelihood that one model is present in a population more frequently than other models being compared. I used the *mfit* package to estimate the protected exceedance probability with the Laplace approximation of the marginal likelihoods of the fitted models. Responses simulated from each model were fit in turn by each of the eight models. As Table 1 shows, each model performed best at fitting the responses generated by its own algorithm in the Decreasing Prevalence condition, with the exception of the Beta-binomial model, suggesting that most of these models would be distinguishable when fitting them to actual human data.

As expected, several models were not reliably distinguishable from one another in the Stable Prevalence condition (see Supplemental Appendix B). This is not surprising, given that Models 2 through 8 are designed to simulate how responses might shift in response to changes in the stimulus distribution, which do not occur in the control condition.

I also used each model to estimate the parameters for each of the twenty simulated participants. The models were reasonably accurate at estimating the true parameters of the simulated agents from that generative model ($r_{mean} = 0.76$), with the exception of the Beta-binomial model ($r = 0.37$). Taken together with the Bayesian model selection for the simulated data, this result suggests that the predicted responses generated by seven of the eight models tested here are distinct, and that approximate recovery of those generative models, as well as estimation of the model parameters for each individual agent, is feasible. Note that the results of the model comparison procedure reported in Section 3.5 below persist whether or not the Beta-binomial model, which performed poorly in model and parameter recovery, is included.

3.4. Parameter estimation for actual human data

Optimized parameters for all 47 human subjects from Study 1 were again estimated using *mfit*. Five uniformly sampled starting values were used for each parameter. The sampling distributions for each parameter were the same as in the simulation and recovery procedure described in

[2] Raising the maximum values of A_0 and B_0 in the Beta-binomial model from 100 to 10,000 did not improve model performance or parameter recovery accuracy.

Table 1
Recovery of simulated data in the Decreasing Prevalence condition of Study 1.

		Model used to fit responses							
		Normal CDF	*MAP*	*Range-Frequency*	*Window*	*AVC*	*Range only*	*Frequency only*	*Beta Binomial*
Model used to generate responses	*Normal CDF*	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	MAP	0.00	**0.99**	0.00	0.00	0.00	0.00	0.00	0.00
	Range-Frequency	0.00	0.00	**0.91**	0.05	0.00	0.01	0.01	0.00
	Window	0.00	0.00	0.02	**0.98**	0.00	0.00	0.00	0.00
	AVC	0.00	0.00	0.00	0.00	**0.99**	0.00	0.00	0.00
	Range only	0.00	0.00	0.01	0.00	0.00	**0.98**	0.00	0.00
	Frequency only	0.01	0.00	0.00	0.00	0.00	0.00	**0.98**	0.00
	Beta Binomial	0.01	0.08	0.02	**0.81**	0.01	0.01	0.03	0.01

Protected exceedance probabilities in model comparison for each fitted model (columns) and the actual generative model of the simulated data (rows) in the Decreasing Prevalence condition. Highest values for each model are bolded.

Section 3.2. For all models, uniform priors were set on each parameter.

3.5. Model comparison for human subjects

I used Bayesian Model Selection as implemented in the *mfit* package in order to compare the models to see which predictions best fit actual human data. The Range-Frequency model strongly outperformed the other models in fitting the data, both across all subjects, $pxp = 0.79$, $BOR < 0.0001$,[3] and within the Decreasing Prevalence condition, $pxp = 0.98$, $BOR < 0.0001$, as shown in Fig. 2. Aggregating BIC values as an alternative form of model comparison produced the same pattern of results.

As expected, the Normal CDF model performed best in the Stable Prevalence condition, $pxp = 0.99$, $BOR < 0.0001$, likely because there was no prevalence shift for these participants to be modeled, and Bayesian Model Selection penalizes other models with added complexity. Fig. 3 shows the choice behavior of each model in the Decreasing Prevalence condition of Study 1. Unlike the Normal CDF model, the Range-Frequency model exhibited a shift in decision thresholds in the decreasing prevalence condition, similar to actual human subjects.

3.6. Discussion

Several mechanisms from cognitive psychology and neuroscience could plausibly explain the phenomenon of prevalence-induced concept change. Here I attempted to use trial-level computational modeling of human data from Study 1 to adjudicate between them. Of the models tested here, Range-Frequency Theory best predicted the judgments of actual human subjects in response to a prevalence decrease.

Can the knowledge that Range-Frequency Theory approximates prevalence-induced concept change inform new interventions to help prevent it? Perhaps the most obvious strategies to do so would involve manipulating either the range parameter, the frequency parameter, or both. Manipulating the frequency parameter in some way to offset a prevalence shift would be difficult, because the very nature of a prevalence shift is that it changes the rank of a given stimulus relative to the local context. Manipulating the range of stimuli seems more feasible, and the high-frequency bursts of "target items" that have been used to compensate for the Low Prevalence Effect in airport baggage screeners (Wolfe, Brunelli, Rubinstein, & Horowitz, 2013) offer a strategy to emulate when designing such interventions. Study 2 was designed to leverage Range-Frequency Theory to reduce or eliminate prevalence-induced concept change in this fashion.

[3] In Bayesian Model Selection, the Bayesian omnibus risk (BOR) reflects the likelihood that the models being compared are all equally frequent in the population.

4. Study 2: Using range extension to eliminate prevalence-induced concept change

4.1. Overview

The computational modeling procedure for Study 1 suggests that Range-Frequency Theory predicts the prevalence-induced concept change exhibited by participants. How could this effect be counteracted? Subjective value of a stimulus in Range-Frequency Theory is determined by a weighted average of the stimulus's intensity compared to the most and least extreme stimuli seen (the "range") and its rank intensity compared to all the other stimuli seen (the "frequency"). In this model, prevalence-induced concept change occurs because decreasing the frequency of stimuli increases the rank of ambiguous stimuli by comparison. It is possible that this effect could be counteracted by increasing the maximum stimulus intensity – using the range extension to offset the frequency decrease. Study 2 was designed to test this prediction and hopefully reduce or eliminate prevalence-induced concept change.

4.2. Methods

4.2.1. Participants

Participants were 110 users of the online survey platform Amazon Mechanical Turk, recruited via CloudResearch.com (59 males, 51 females, $M_{age} = 37.66$ years, $SD = 11.27$ years) who were paid $2 USD for their participation.

4.2.2. Procedure

As in Study 1, participants were told that a series of colored dots would appear on the screen, one at a time, and that their task was to decide whether each dot was blue or not blue, and to indicate their decision by pressing one of two keys on the keyboard ("f" for not blue, "j" for blue).

On each trial, a colored dot appeared on a solid gray background. The color of the dot varied across trials from very purple (61% blue, RGB 98–0-156) to very blue (100% blue, RGB 0–0-254). Each dot appeared on the screen for 500 milliseconds and was then replaced by a question mark, which remained on the screen until participants pressed one of the response keys. Participants were told that there would be 400 trials divided into 8 blocks, and that the prevalence of blue dots might vary across blocks. Specifically, they were told that some blocks "may have a lot of blue dots, and others may have only a few." Participants completed 10 practice trials to ensure they understood the procedure, and then completed 400 test trials. To help participants remain attentive, I allowed them to take a brief break every 50 trials.

Participants were randomly assigned to one of four conditions in a 2 (prevalence: *stable* or *decreasing*) x 2 (range: *fixed* or *extending*) design. As in Study 1, I divided the dots participants could see into two spectra. The prevalence condition determined what proportion of dots participants saw from each of those sections. Participants in the *stable prevalence* conditions saw dots that had an equal probability (50%) of being drawn

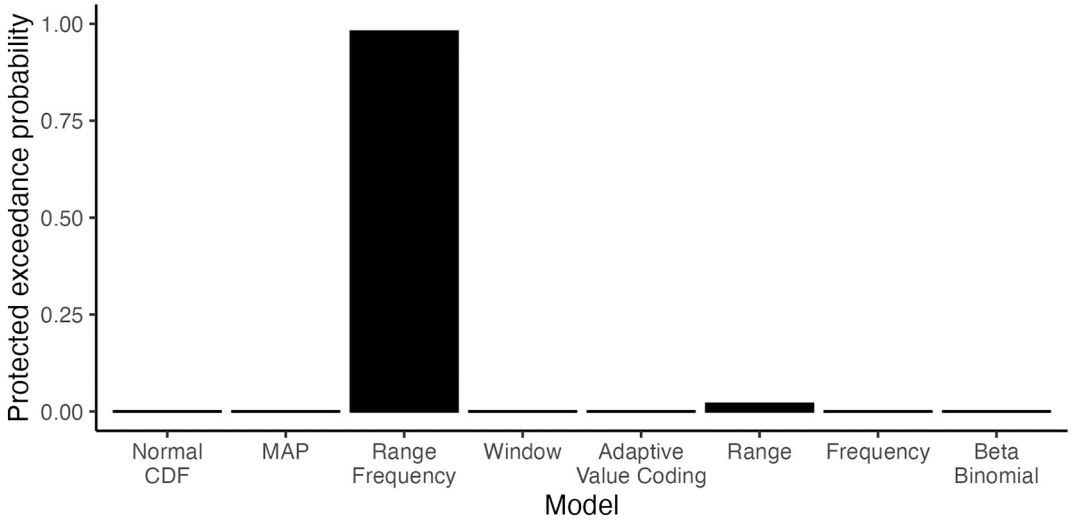


Fig. 2. Bayesian Model Selection of Human Data in the Decreasing Prevalence Condition of Study 1.
Protected exceedance probability (PXP) scores (y-axis) are shown for color identification data from the subjects in the Decreasing Prevalence Condition of Study 1, fit by eight different models (x-axis).

from the "blue" or "purple" sections of the color spectrum on all 400 trials. Participants in the *decreasing prevalence* conditions also saw a 50% prevalence of blue dots for the first 200 trials, but for trials 201–400, instead only saw a 10% prevalence of blue dots.

For the first 200 trials, all participants saw dots from a "narrow" blue spectrum, RGB 78–0-176 (69% blue) through RGB 50–0-204 (80% blue). The range condition determined the upper range of colors participants saw in the final 200 trials of the study. Starting on trial 201, participants in the *extending range* conditions saw dots from a "wide" blue spectrum, from RGB 78–0-176 (69% blue) through RGB 0–0-254 (100% blue). Participants in the *fixed range* conditions instead continued to see dots from the "narrow" blue spectrum. After completing the task, participants completed a questionnaire asking some basic demographics and their impressions of the task. The complete text of the task instructions and all questions is available in Supplemental Appendix A.

4.3. Results

Following my preregistration, I excluded eight participants whose individual responses could not be fit to a gaussian CDF with free parameters for threshold and sensitivity, suggesting inattention or random button pressing. Importantly, the results reported in this section do not change whether or not these participants are included in the analysis.

Did extending the range of colors reduce the typical effect of the prevalence decrease? To find out, I fit a binomial generalized linear mixed model to my data in R using the package *lme4*. The dependent variable was the participant's *identification* of a dot as blue or not blue. The independent between-participants variables were the participant's assigned *prevalence* (stable or decreasing) and *range* (fixed or extending). The independent within-participants variables were (a) the dot's RGB value or what I will call its *actual color* (which ranged from 61% blue to 100% blue, coded as 0-100) and (b) the *trial number* (which ranged from 1 to 400). I also included interactions between all fixed effects in my model. I included a random intercept term for participants (who may have entered my study with different thresholds) and allowed slopes to vary randomly by trial number for each participant. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 1395.20$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 590.95$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and actual color significantly improved model fit, $\chi^2(1) = 14.61$, $p < 0.001$.

The generalized linear mixed model revealed that a Condition X Actual Color X Trial Number X Range interaction predicted participants' identifications, $b = 14.01$, $SE = 0.37$, $z = 37.96$, $p < 0.001$, *95% CI*$_b$ [13.29, 14.74], $R^2_{GLMM(c)} = 0.86$. Fig. 4 shows the percentage of dots at each point along the continuum that participants identified as blue on the initial 200 trials and on the final 200 trials. When the range of stimuli was fixed, shown in the two left panels, participants only called a wider range of colors blue in the final trials when the prevalence of blue dots decreased over time, replicating the results of Study 1. When the range of blue colors participants saw was extended over time but the prevalence of blue stayed the same, shown in the upper right panel, participants instead called a narrower range of colors blue in the final trials compared to the earlier trials. Crucially, when the range of colors was extended over time and the prevalence of blue decreased, as in the lower right panel, participants did not call a wider or narrower range of colors blue in the final trials compared to the initial trials. In other words, extending the range of stimuli eliminated the prevalence decrease's effect on judgments and caused participants to maintain their original color thresholds, even when blue dots became rare. In Supplemental Appendix C, I report a computational modeling procedure of these data similar to that performed for Study 1, which suggests that the same Range-Frequency Model which best described the results of Study 1 can also do so for Study 2.

5. General discussion

In this paper, I have attempted to provide evidence for a computational mechanism that can explain prevalence-induced concept change. In Study 1, decreasing the prevalence of blue dots caused observers to call a wider range of colors blue, replicating the findings of Levari et al. (2018). A comparison of several plausible computational models suggested that the process that best characterized the results was Range-Frequency Theory, a weighted compromise between the frequency and range of recently seen stimuli. Finally, in Study 2, an intervention designed with this mechanism in mind successfully eliminated prevalence-induced concept change by increasing the range of stimuli as they became rare.

5.1. Limitations and future directions

Many of the conclusions about the meaning and implications of these findings rest on the assumption, supported by the computational models presented here, that Range-Frequency Theory is a good candidate

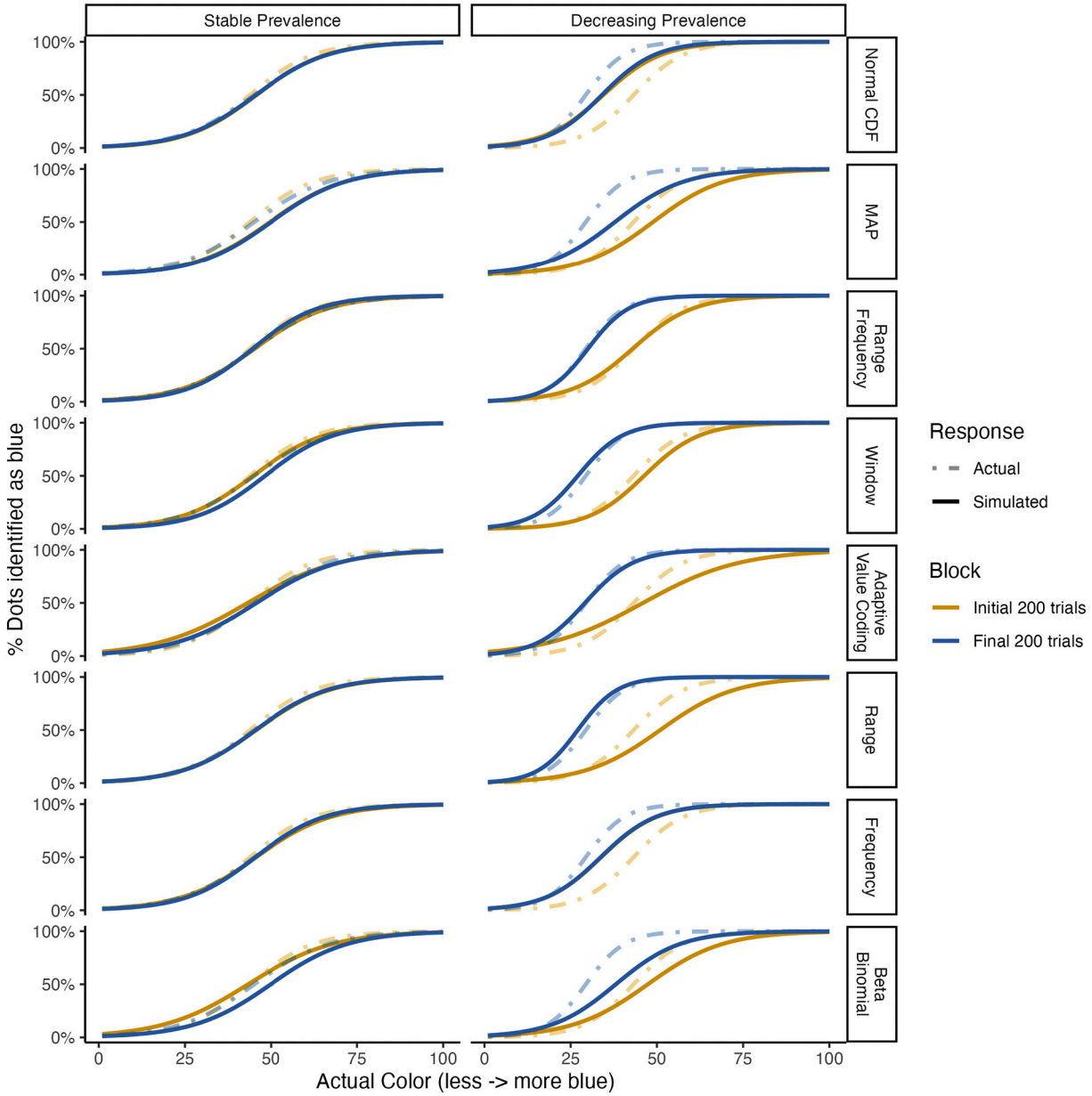


Fig. 3. Choice behavior of fitted models on the color identification task in Study 1.
The x-axes show the dot's objective color and the y-axes show the percentage of trials on which simulated agents from each model (solid lines) or actual human participants from Study 1 (dashed lines) identified that color as blue. Fitted lines were computed as binomial GLMs.

mechanism to describe how humans implement prevalence-induced concept change on a cognitive level. However, in this paper I have only explored a small subset of the possible mechanisms that could predict similar effects, and only for one kind of task, color identification. Many sophisticated models of contextual perception and valuation exist in the cognitive sciences. For example, in Haubensak (1992), judgments are made using a scale centered on stimulus values presented early on in a sequence, which can produce judgment effects that appear driven by frequency, simply because the initial values will also typically be the most frequent over time. In Decision by Sampling (Stewart, Chater, & Brown, 2006), no explicit value of stimuli is directly computed, and valuations come from comparisons with local context. It is possible that such a model or another model entirely (e.g. Wilson, 2018) would better predict the findings presented here, or previously documented prevalence effects in other domains such as moral judgment. Bhui and Gershman (2018) have recently argued that Decision by Sampling and Range-Frequency Theory are in fact closely related formulations of the principle of efficient coding in valuation. This convergence of existing theories of contextual effects on judgment lends support to the possibility that prevalence-induced concept change is driven by some implementation of efficient coding in neural architecture. Hopefully, the findings presented here can help inform future research to further refine our understanding of the computational and neural processes that produce contextual effects in many different domains, particularly in cases like this one, where multiple theories plausibly explain the findings in question.

5.2. *Concluding remarks*

As Levari et al. (2018) note, the fact that nominally fixed concepts are susceptible to prevalence-induced change may have troubling consequences. One such consequence is that decision makers may find it difficult to tell when undesirable things are in fact becoming less common over time. Governments, institutions, and organizations seek to

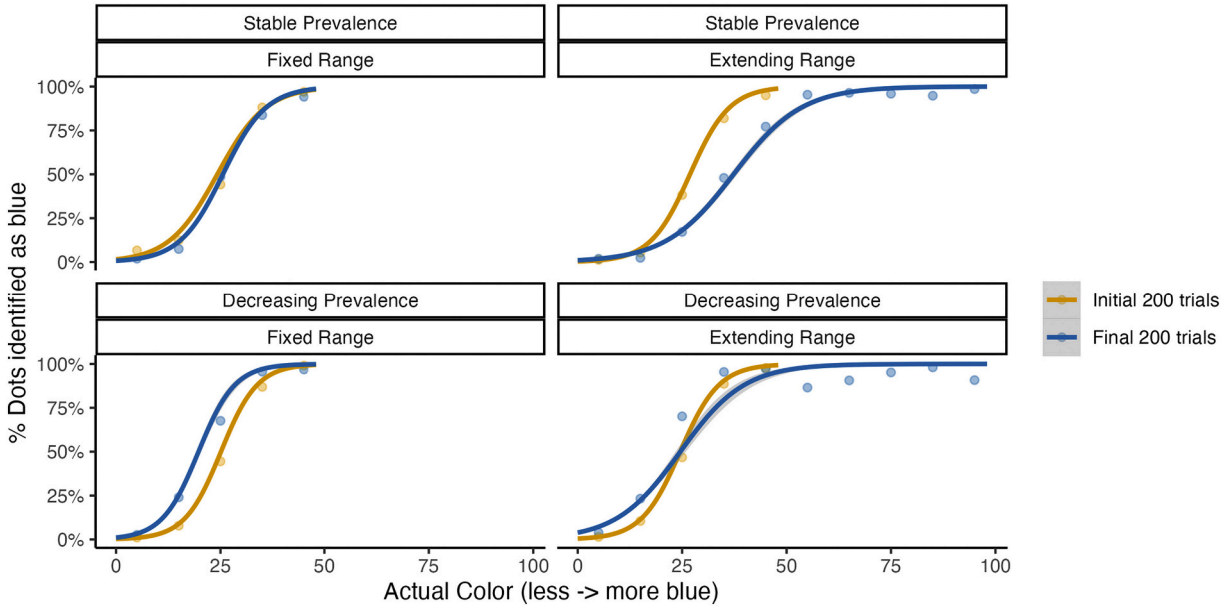


Fig. 4. Results of Study 2.
The x axis shows the dot's objective color, and the y axis shows the percentage of trials on which participants identified that dot as blue. Fitted lines were computed as binomial GLMs.

identify problems so that they can then take action to decrease their future prevalence. IRB reviewers, for example, seek to identify unethical research projects not only to keep them from being executed, but also to reduce the number of unethical projects that researchers propose in the future. When researchers respond precisely as reviewers hope, reviewers may unwittingly expand their concepts of unethicality and start rejecting proposals that they would earlier have accepted, effectively "moving the goalposts" of their ethical standard. This phenomenon is not limited to IRBs, and may plague the well-meaning enforcers of many policies that require people to ameliorate the thing they are attempting to assess.

Another possible consequence is that observers may quickly become desensitized to problems when they proliferate. For example, if a business is attempting to reduce corruption or other unethical practices, any dramatic increase in those practices could lead to relaxed standards for what counts as an ethical violation. In this case, the simple fact of unethical behavior growing more prevalent would lead to more ethical lapses being forgiven, a well-known finding in research on contextual effects in ethical and unethical behavior (Aldrovandi, Wood, & Brown, 2013; Gino & Bazerman, 2009; Marsh & Parducci, 1978).

How could an understanding of Range-Frequency theory help keep judgments consistent in domains where changes in prevalence can wreak havoc on standards? Perhaps the simplest application of the results of Study 2 would be to give people referents to keep in mind that are tailored to the current prevalence of the stimuli they need to judge. For example, an IRB officer could be periodically shown some examples of extremely unethical studies for comparison when the prevalence of unethical study proposals is low. If the prevalence increases, they could then instead compare submissions to examples of moderately unethical studies. This approach has the disadvantage of requiring real-time adjustment of referents based on changing prevalence, which institutions may be unable to precisely track over time. Future work should investigate whether consistently maintaining a fixed or variable range of comparators can preemptively prevent prevalence from influencing judgments in undesirable ways. Hopefully, range extension can be added to the presence or absence of feedback (Lyu, Levari, Nartker, Little, & Wolfe, 2021) as a useful tool for policymakers and practitioners who wish to keep evaluations consistent over time, even in the face of prevalence shifts.

Author contributions

D. Levari conceptualized, designed, and analyzed the studies, collected the data, and wrote the paper.

Declaration of Competing Interest

None.

Acknowledgements

I thank Adam Bear, Rahul Bhui, Sean Devine, Daniel Gilbert, Sam Gershman, Joshua Greene, Adam Morris, and Jeremy Wolfe for helpful comments, Steven Worthington of the Institute for Quantitative Social Science at Harvard University for statistical support, and John Amhanesi, Zach Diamandis, Jillian Graver, Gretchen Shoenberger, Amaya Sizer, and Jarod Stone for research assistance. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary Data and Information

Data, materials, code for analysis, and preregistrations are available at the following link: https://osf.io/b24tv/. Supplementary information for this article can be found online at https://doi.org/10.1016/j.cognition.2022.105196.

References

Aldrovandi, S., Wood, A. M., & Brown, G. D. A. (2013). Sentencing, severity, and social norms: A rank-based model of contextual influence on judgments of crimes and punishments. *Acta Psychologica, 144*(3), 538–547. https://doi.org/10.1016/j.actpsy.2013.09.007

Anstis, S., Verstraten, F. A. J., & Mather, G. (1998). The motion aftereffect. *Trends in Cognitive Sciences, 2*(3), 111–117. https://doi.org/10.1016/S1364-6613(98)01142-5

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183–193. https://doi.org/10.1037/h0054663

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems, 12*(3), 241–253. https://doi.org/10.1088/0954-898X/12/3/301

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v067.i01

Bhui, R., & Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review, 125*(6), 985–1001. https://doi.org/10.1037/rev0000123

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences, 41*, 15–21. https://doi.org/10.1016/j.cobeha.2021.02.015

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *The American Journal of Psychology, 96*(1), 85–94. https://doi.org/10.2307/1422211

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *The Journal of Neuroscience, 17*(21), 8621–8644.

Cicchini, G. M., Mikellidou, K., & Burr, D. (2017). Serial dependencies act directly on perception. *Journal of Vision, 17*(14). https://doi.org/10.1167/17.14.6

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In *, 23. Decision Making, Affect, and Learning: Attention and Performance XXIII* (pp. 3–38). https://doi.org/10.1093/acprof

Decarlo, L. T. (2013). Signal detection models for the same – different task. *Journal of Mathematical Psychology, 57*(1–2), 43–51. https://doi.org/10.1016/j.jmp.2013.02.002

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2018). Due process in dual process: Model-recovery simulations of decision-bound strategy analysis in category learning. *Cognitive Science, 42*(S3), 833–860. https://doi.org/10.1111/cogs.12607

Feldman, J. (2021). Information-theoretic signal detection theory. *Psychological Review, 128*(5), 976–987. https://doi.org/10.1037/rev0000300

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116*(4), 752–782. https://doi.org/10.1037/a0017196

Fornaciai, M., & Park, J. (2018). Attractive serial dependence in the absence of an explicit task. *Psychological Science, 29*(3), 437–446. https://doi.org/10.1177/0956797617737385

Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite effects of recent history on perception and decision. *Current Biology, 27*(4), 590–595. https://doi.org/10.1016/j.cub.2017.01.006

Gershman, S. (2015). *MFIT* [MATLAB]. https://github.com/sjgershm/mfit.

Gino, F., & Bazerman, M. H. (2009). When misconduct goes unnoticed: The acceptability of gradual erosion in others' unethical behavior. *Journal of Experimental Social Psychology, 45*(4), 708–719. https://doi.org/10.1016/j.jesp.2009.03.013

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.12504. n/a–n/a.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 1–49). Cambridge University Press.

Griffiths, T. L., & Yuille, A. (2008). A primer on probabilistic inference. In M. Oaksford, & N. Chater (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science*. https://doi.org/10.1093/acprof:oso/9780199216093.003.0002

Haubensak, G. (1992). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance, 18*(1), 303–309. https://doi.org/10.1037/0096-1523.18.1.303

Hayden, B. Y. (2018). Economic choice: The foraging perspective. *Current Opinion in Behavioral Sciences*. https://doi.org/10.1016/j.cobeha.2017.12.002

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*(2), 181–197. https://doi.org/10.1017/S0952523800009640

Helson, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior*. Harper & Row.

Khaw, M. W., Glimcher, P. W., & Louie, K. (2017). Normalized value coding explains dynamic adaptation in the human valuation process. *Proceedings of the National Academy of Sciences, 201715293*. https://doi.org/10.1073/pnas.1715293114

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods, 47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. *Science, 360*(6396), 1465–1467. https://doi.org/10.1126/science.aap8731

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), 6139–6144. https://doi.org/10.1073/pnas.1217854110

Lyu, W., Levari, D. E., Nartker, M. S., Little, D. S., & Wolfe, J. M. (2021). Feedback moderates the effect of prevalence on perceptual decisions. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-01956-3

Manassi, M., Liberman, A., Kosovicheva, A., Zhang, K., & Whitney, D. (2018). Serial dependence in position occurs at the time of perception. *Psychonomic Bulletin & Review, 25*(6), 2245–2253. https://doi.org/10.3758/s13423-018-1454-5

Marsh, H. W., & Parducci, A. (1978). Natural anchoring at the neutral point of category rating scales. *Journal of Experimental Social Psychology, 14*(2), 193–204. https://doi.org/10.1016/0022-1031(78)90025-2

Mather, G., Pavan, A., Campana, G., & Casco, C. (2008). The motion aftereffect reloaded. *Trends in Cognitive Sciences, 12*(12), 481–487. https://doi.org/10.1016/j.tics.2008.09.002

McNair, J. N. (1982). Optimal giving-up times and the marginal value theorem. *The American Naturalist, 119*(4), 511–529. http://www.jstor.org/stable/10.2307/2461141.

Paradiso, M. A., Shimojo, S. S., & Nakayama, K. E. N. (1989). Subjective contours, tilt aftereffects, and visual cortical organization. *Vision Research, 29*(9), 1205–1213. https://doi.org/10.1016/0042-6989(89)90066-7

Parducci, A. (1963). Range-frequency compromise in judgment. *Psychological Monographs: General and Applied, 77*(2), 1–50. https://doi.org/10.1037/h0093829

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*(6), 407–418. https://doi.org/10.1037/h0022602

Parducci, A., & Wedell, D. H. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance, 12*(4), 496–516. https://doi.org/10.1037/0096-1523.12.4.496

R Core Team. (2020). R Core team, R: A language and environment for statistical computing. http://www.r-project.org/.

Rhodes, G., Jeffery, L., Watson, T. L., Clifford, C. W. G., & Nakayama, K. (2003). Fitting the mind to the world: Face adaptation and attractiveness aftereffects. *Psychological Science, 14*(6), 558–566. https://doi.org/10.1046/j.0956-7976.2003.psci_1465.x

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—Revisited. *NeuroImage, 84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Schwartz, O., Hsu, A., & Dayan, P. (2007). Space and time in visual context. *Nature Reviews. Neuroscience, 8*(7), 522–535. https://doi.org/10.1038/nrn2155

Spektor, M. S., Bhatia, S., & Gluth, S. (2021). The elusiveness of context effects in decision making—ClinicalKey. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2021.07.011

Stankevicius, A., Huys, Q. J. M., Kalra, A., & Seriès, P. (2014). Optimism as a prior belief about the probability of future reward. *PLoS Computational Biology, 10*(5), Article e1003605. https://doi.org/10.1371/journal.pcbi.1003605

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience, 1–12*. https://doi.org/10.1038/nrn3838

Vickers, D., & Leary, J. N. (1983). Criterion control in signal detection. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 25*(3), 283–296. https://doi.org/10.1177/001872088302500305

Webb, R., Glimcher, P. W., & Louie, K. (2020). Divisive normalization does influence decisions with multiple alternatives. *Nature Human Behaviour, 1–3*. https://doi.org/10.1038/s41562-020-00941-5

Webster, M. (1996). Human colour perception and its adaptation. *Network: Computation in Neural Systems, 7*(4), 587–634. https://doi.org/10.1088/0954-898x/7/4/002

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience, 18*(10), 1509–1517. https://doi.org/10.1038/nn.4105

Wilson, R. C. (2018). Sequential choice effects predict prevalence-induced concept change. *PsyArXiv.*. https://doi.org/10.31234/osf.io/75bpy

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife, 8*, Article e49547. https://doi.org/10.7554/eLife.49547

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *Journal of Vision, 13*(3), 10. https://doi.org/10.1167/13.3.10

Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision, 13*(3), 33. https://doi.org/10.1167/13.3.33

Supplemental Appendix

For

Range-frequency effects can explain and eliminate prevalence-induced concept change

by

David E. Levari

Table of Contents

(Appendix A) Instructions and Post-Task Questions for Studies 1 and 2

Study 1 Instructions

Welcome to this study! We're interested in studying how people perceive and identify colors.
In this task, you will see dots presented on the screen one at a time, in a variety of colors. Your task in this study will be to identify blue dots.
When you see a blue dot on the screen, press the "blue" key. For all other dots, press the "not blue" key.
The dots will be presented in series with breaks in between. This means that you will see a series of dots, have a short break, and then another series of dots, until you have seen 16 series.
Some of the series you see may have a lot of blue dots, and other may have only a few. There's nothing for you to count or keep track of -- your only task is to identify blue dots.
You should do your best to answer quickly and accurately during the study. However, if you make a mistake and hit the wrong button at any point, just keep going.
Now you will complete a brief practice series so you can get used to the task.
You have now completed the practice series. If you have any questions, you can ask the experimenter now.
Otherwise, you're ready to begin the study.

Study 1 Post-task questionnaire

Thanks for participating in the study! Please answer a few last questions before you go.

How old are you?

Please indicate your gender:
- ❍ Male (1)
- ❍ Female (2)
- ❍ Prefer not to answer (3)

Did you find the task easy or difficult?
- ❍ Very easy (1)
- ❍ Easy (2)
- ❍ Somewhat Easy (3)
- ❍ Neutral (4)
- ❍ Somewhat Difficult (5)
- ❍ Difficult (6)
- ❍ Very Difficult (7)

Are you right or left handed?

- ❍ Right handed (1)
- ❍ Left handed (2)

Do you wear corrective lenses? If so, are you wearing them right now?

- ❍ Yes, but I'm not wearing them now (1)
- ❍ Yes, and I am wearing them now (2)
- ❍ No, I'm don't wear corrective lenses (3)

Is English your only native language?

- ❍ Yes (1)
- ❍ No, English is not my native language (2)
- ❍ No, I spoke English and other languages growing up (3)

What do you think this study was about?

Do you think that it became easier or harder to find blue dots as the study progressed?

- ❍ It became easier to find blue dots as the study progressed (1)
- ❍ It became harder to find blue dots as the study progressed (2)
- ❍ It was about the same throughout the study (3)
- ❍ I'm not sure (4)

Do you feel that the amount of blue dots in each series changed during the study?

- ❍ No, there were the same number of blue dots throughout the study (1)
- ❍ Yes, there were fewer blue dots as the study went on (2)
- ❍ Yes, there were more blue dots as the study went on (3)
- ❍ I'm not sure (4)

We want to get a sense of how many blue dots you think you saw at different times in the study. Please indicate, using the options below, your impressions about what proportion of the dots you saw were blue. If you have no idea, please check the box labeled "Not sure" instead of using the slider.

_____ In the first few series, I saw... (1)
_____ The the middle few series, I saw... (2)
_____ In the last few series, I saw... (3)

By the end of the study, do you think that your definition of what counted as a "blue" dot changed?

- ❍ No, I think that my definition of what counted as a blue dot did not change during the study. (1)
- ❍ Yes, I think my definition of what counts as a blue dot expanded -- I counted a wider range

of colors as blue at the end of the study compared to the beginning of the study. (2)

- ❍ Yes, I think my definition of what counts as a blue dot narrowed -- I counted a smaller range of colors as blue at the end of the study compared to the beginning of the study. (3)
- ❍ I'm not sure if my definition changed (4)
- ❍ I don't understand this question (5)

If you have any other comments about the study, please let us know here:

Study 2 Instructions

Welcome to this study! We're interested in studying how people perceive and identify colors.
In this task, you will see dots presented on the screen one at a time, in a variety of colors. Your task in this study will be to identify blue dots.
When you see a blue dot on the screen, press the "blue" key. For all other dots, press the "not blue" key.
The dots will be presented in series with breaks in between. This means that you will see a series of dots, have a short break, and then another series of dots, until you have seen 8 series.
Some of the series you see may have a lot of blue dots, and other may have only a few. There's nothing for you to count or keep track of -- your only task is to identify blue dots.
You should do your best to answer quickly and accurately during the study. However, if you make a mistake and hit the wrong button at any point, just keep going.
Now you will complete a brief practice series so you can get used to the task.
You have now completed the practice series. If you have any questions, you can ask the experimenter now.
Otherwise, you're ready to begin the study.

Study 2 Post-task questionnaire

Thanks for participating in the study! Please answer a few last questions before you go.

How old are you?

Please indicate your gender:

- ❍ Male (1)
- ❍ Female (2)
- ❍ Prefer not to answer (3)

Did you find the task easy or difficult?

- ❍ Very easy (1)
- ❍ Easy (2)
- ❍ Somewhat Easy (3)
- ❍ Neutral (4)
- ❍ Somewhat Difficult (5)
- ❍ Difficult (6)
- ❍ Very Difficult (7)

Are you right or left handed?

- ❍ Right handed (1)
- ❍ Left handed (2)

Do you wear corrective lenses? If so, are you wearing them right now?

- ❍ Yes, but I'm not wearing them now (1)
- ❍ Yes, and I am wearing them now (2)
- ❍ No, I'm don't wear corrective lenses (3)

Is English your only native language?

- ❍ Yes (1)
- ❍ No, English is not my native language (2)
- ❍ No, I spoke English and other languages growing up (3)

What do you think this study was about?

Do you think that it became easier or harder to find blue dots as the study progressed?

- ❍ It became easier to find blue dots as the study progressed (1)
- ❍ It became harder to find blue dots as the study progressed (2)
- ❍ It was about the same throughout the study (3)
- ❍ I'm not sure (4)

Do you feel that the amount of blue dots in each series changed during the study?

- ❍ No, there were the same number of blue dots throughout the study (1)
- ❍ Yes, there were fewer blue dots as the study went on (2)
- ❍ Yes, there were more blue dots as the study went on (3)
- ❍ I'm not sure (4)

We want to get a sense of how many blue dots you think you saw at different times in the study. Please indicate, using the options below, your impressions about what proportion of the dots you

saw were blue. If you have no idea, please check the box labeled "Not sure" instead of using the slider.

_____ In the first few series, I saw... (1)
_____ The the middle few series, I saw... (2)
_____ In the last few series, I saw... (3)

Do you feel that the range of colors in each series changed during the study?"

- ❍ No, there was the same range of colors throughout the study (1)
- ❍ Yes, there were was a wider range of colors as the study went on (2)
- ❍ Yes, there was a narrower range of colors as the study went on (3)
- ❍ I'm not sure (4)

By the end of the study, do you think that your definition of what counted as a "blue" dot changed?

- ❍ No, I think that my definition of what counted as a blue dot did not change during the study. (1)
- ❍ Yes, I think my definition of what counts as a blue dot expanded -- I counted a wider range of colors as blue at the end of the study compared to the beginning of the study. (2)
- ❍ Yes, I think my definition of what counts as a blue dot narrowed -- I counted a smaller range of colors as blue at the end of the study compared to the beginning of the study. (3)
- ❍ I'm not sure if my definition changed (4)
- ❍ I don't understand this question (5)

If you have any other comments about the study, please let us know here:

(Appendix B) Recovery of simulated data in the control condition of Study 1

Table B.1: Recovery of simulated data in the control condition of Study 1

		Model used to fit responses							
		Normal CDF	*MAP*	*Range-Frequency*	*Window*	*AVC*	*Range only*	*Frequency only*	*Beta Binomial*
Model used to generate responses	*Normal CDF*	0.02	0.02	0.35	0.02	0.02	0.15	**0.39**	0.02
	MAP	0.00	0.00	0.00	**0.99**	0.00	0.00	0.00	0.00
	Range-Frequency	0.03	0.02	**0.81**	0.06	0.02	0.02	0.02	0.02
	Window	**0.76**	0.07	0.02	0.07	0.02	0.02	0.02	0.02
	AVC	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00
	Range only	**0.99**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Frequency only	0.11	0.09	**0.26**	0.12	0.09	0.13	0.11	0.09
	Beta Binomial	0.05	0.05	0.17	**0.50**	0.05	0.07	0.07	0.05

Protected exceedance probabilities in model comparison for each fitted model (columns) and the actual generative model of the simulated data (rows). Highest values for each model are bolded.

(Appendix C) Computational modeling of Study 2 data

Here I report a computational modelling analysis for Study 2, in which the Range-Frequency Model which best predicted the data in Study 1 is compared to a control model to see which best predicts the behavioral results from Study 2. I first use both models to simulate data in all four conditions in the color identification task in Study 2, and confirm that each model can not only accurately estimate the true simulation starting parameters, but also recover its own

simulated responses better than the other model. Then, I fit both models to the data from Study 2, using Bayesian Model Selection to determine which model best accounts for actual human responses in the task.

Model specifications

The models used are Models 1 (Normal CDF) and Model 3 (Range-Frequency) as described in the main text.

Generation of simulated data

Data for 100 simulated agents in the color identification task from Study 2 (25 in each condition) were generated in MATLAB from each of the two models being tested. The true parameters for each agent were randomly sampled from uniform distributions. The parameter specifications were the same as those described in the main text. Each agent completed 400 trials. I then used a trial-level model fitting procedure with maximum likelihood estimation to recover the true generative model and parameters of the simulated data.

Recovery of simulated data

Optimized parameters for each simulated agent were estimated with the MATLAB package *mfit* (Gershman, 2015/2021). Five starting values were uniformly sampled for each parameter, with the same bounds as used for data generation (described above).

To test the ability of each model to recover its own simulated responses, I calculated a protected exceedance probability for the simulated responses from each of the two models, fit in turn by each of the two models (a total of 4 model fits). I used the *mfit* package to estimate the protected exceedance probability with the Laplace approximation of the marginal likelihoods of the fitted models. Responses simulated from each model were fit in turn by each of the two models. As Table C.1 shows, each model performed best at fitting the responses generated by its

own algorithm, suggesting that these models would be distinguishable when fitting them to actual human data.

Table C.1: Recovery of simulated data in Study 2

		Model used to fit responses	
		Normal CDF	*Range-Frequency*
Model used to generate responses	*Normal CDF*	**0.98**	0.02
	Range-Frequency	0.00	**1.00**

Protected exceedance probabilities in model comparison for each fitted model (columns) and the actual generative model of the simulated data (rows). Highest values for each model are bolded.

I also used each model to estimate the parameters for each of the 100 simulated participants. Both models were reasonably accurate at estimating the true parameters of the simulated agents from that generative model ($r_{mean} = 0.85$). Taken together with the Bayesian model selection for the simulated data, this result suggests that the predicted responses generated by the two models tested here are distinct, and that approximate recovery of those generative models, as well as estimation of the model parameters for each individual agent, is feasible.

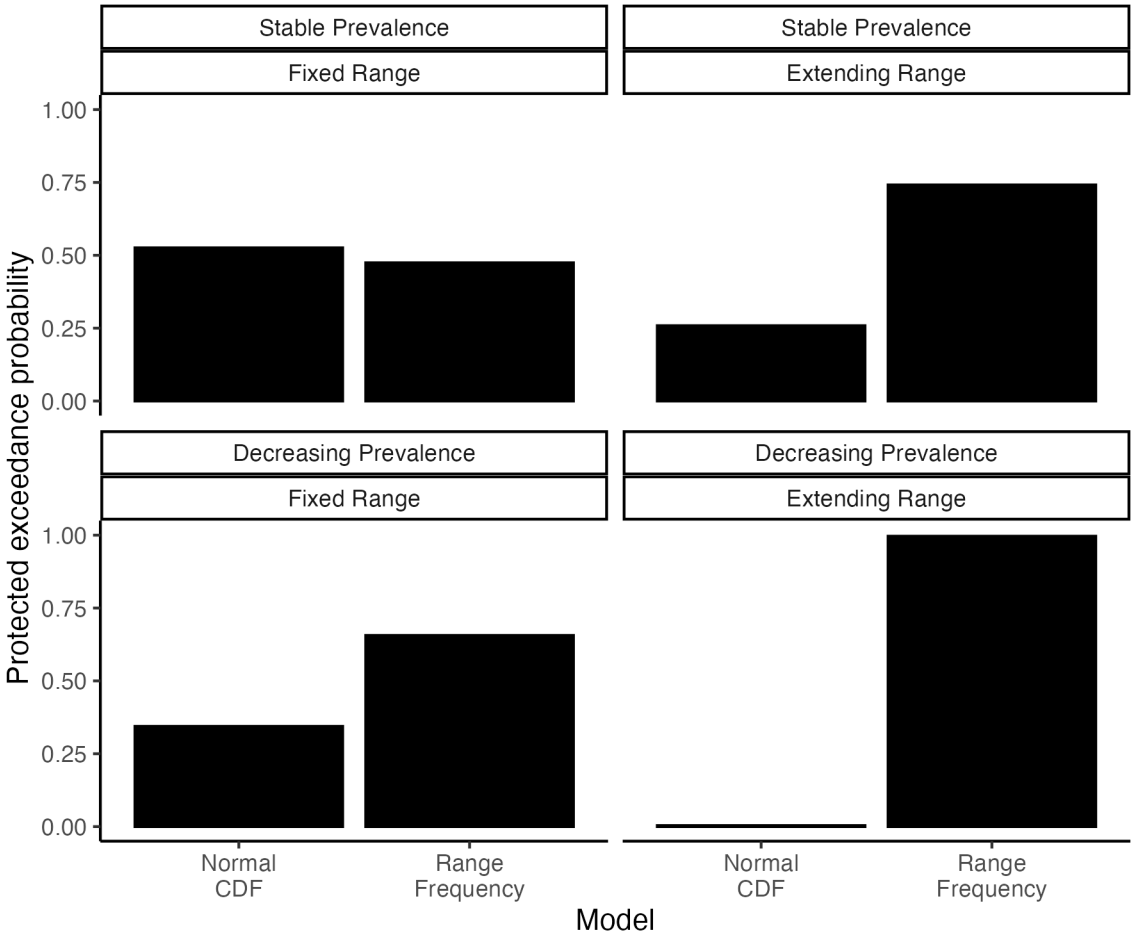
Parameter estimation for actual human data

Optimized parameters for all 102 human subjects from Study 2 were again estimated using *mfit*. Five uniformly sampled starting values were used for each parameter. The sampling distributions for each parameter were the same as in the simulation and recovery procedure described above. For all models, uniform priors were set on each parameter.

Model comparison for human subjects

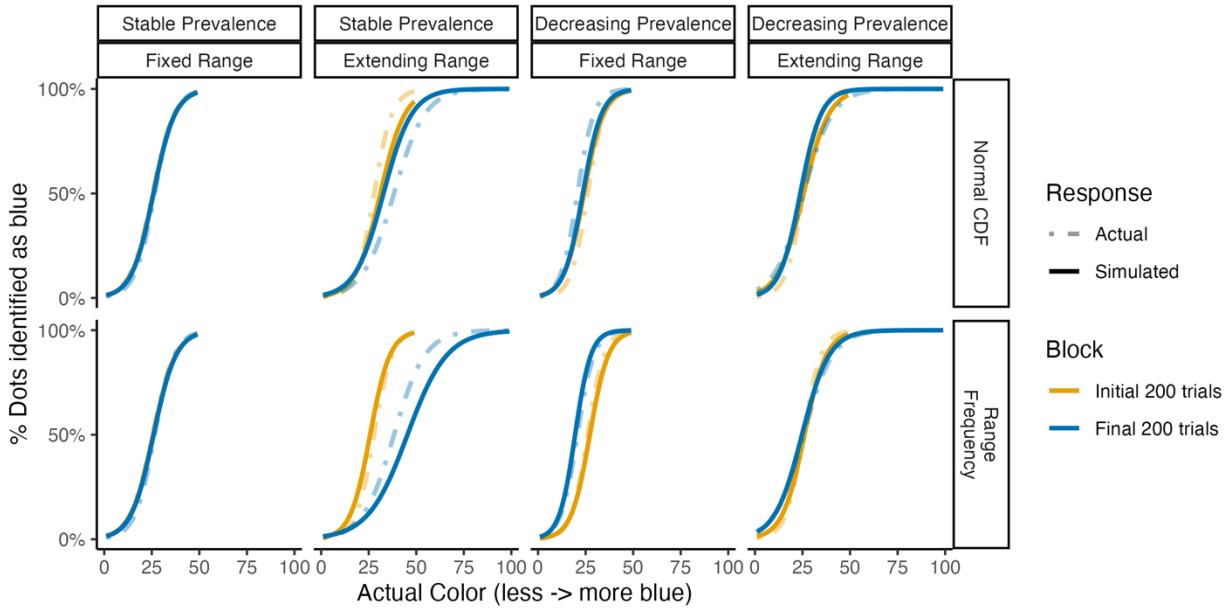
I used Bayesian Model Selection as implemented in the *mfit* package in order to compare the models to see which predictions best fit actual human data. Figure C.1 shows the protected exceedance probabilities of the eight models. The Range-Frequency model outperformed the Normal CDF model in fitting the data across all subjects ($pxp = 0.99$, $BOR = 0.02$), and in the three conditions in which prevalence and/or range were manipulated: the *Stable Prevalence/Extending Range* condition ($pxp = 0.74$), the *Decreasing Prevalence/Fixed Range* condition ($pxp = 0.66$), and the *Decreasing Prevalence/Extending Range* condition ($pxp = 0.99$). Aggregating BIC values as an alternative form of model comparison produced the same pattern of results. Figure C.2 shows the choice behavior of each model in each condition of Study 2. Unlike the Normal CDF model, simulated responses from the Range-Frequency model showed a shift in decision thresholds during the task in the *Stable Prevalence/Extending Range* condition and the *Decreasing Prevalence/Fixed Range* condition, similar to what was observed in actual human subjects.

Figure C.1: Bayesian Model Selection of Human Data in Study 2



Protected exceedance probability (PXP) scores (y-axis) are shown for color identification data from the subjects each condition of Study 2, fit by two different models (x-axis).

Figure C.2: Choice behavior of fitted models on the color identification task in Study 2



The x-axes show the dot's objective color (i.e., its location on the continuum) and the y-axes show the percentage of trials on which simulated agents from each model (solid lines) or actual human participants from Study 2 (dashed lines) identified that color as blue. Fitted lines were computed as binomial GLMs.
