Prevalence-Induced Concept Change in Human Judgment

Prevalence-Induced Concept Change in Human Judgment

A dissertation presented

by

David Emmanuel Levari

to

The Department of Psychology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Psychology

Harvard University

Cambridge, Massachusetts

May, 2018

Dissertation Advisor: Professor Daniel T. Gilbert David Emmanuel Levari

Prevalence-Induced Concept Change in Human Judgment

ABSTRACT

Why do social problems seem so intractable? In a series of experiments, I show that people often respond to decreases in the prevalence of a stimulus by "expanding their concept" of it. When blue dots became rare, participants began to see purple dots as blue; when threatening faces became rare, participants began to see neutral faces as threatening; and when unethical requests became rare, participants began to see innocuous requests as unethical. This *prevalence-induced concept change* occurred even when participants were forewarned about it and even when they were instructed and paid to resist it. I then present a computational model suggesting that this phenomenon is driven by a range-frequency compromise in judgment. One reason social problems may seem so intractable is that a reduction in their prevalence can lead people to see more of them.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

To my doctoral advisor, Daniel Gilbert. I won the lottery when I got a chance to join his lab. He's an incomparable mentor and colleague, and I never imagined that I could learn so much from his guidance while always feeling like a collaborator as well as a student. This project is as much his as mine, and I'll treasure the thrill that we shared as we discovered something interesting and unexpected about those old strange machines, human minds.

Many of the seeds of this project originated in conversations Dan had with Richard Hackman, many years ago. Richard passed away in 2013, just as we started running these experiments. I was lucky to get to know him, all too briefly. I like to think that he would have gotten a kick out of these results, and I'm grateful that his lifelong curiosity and knack for finding fascinating questions helped spur all of this work.

My dissertation committee members, Joshua Greene, Jeremy Wolfe, and Sam Gershman, were incredibly helpful and generous with their time and wisdom, both in guiding where this project went, and giving their candid thoughts on how it turned out. If every graduate student had a committee as helpful and kind as these three, the process of writing a dissertation would have a reputation for being a pleasure, rather than a slog.

To the many kind people I received technical and statistical help from while working on these studies: Tim Brady, Adam Bear, Steven Worthington and Ista Zahn from the Harvard Institute for Quantitative Social Science (IQSS), Patrick Mair, and Mark Thornton. Adam Morris spent many painstaking hours teaching me almost everything I know about computational modeling, which is still no doubt only a small sliver of what he knows.

To my labmates, collaborators, and other colleagues who were always supportive and saw, read, and heard way too many drafts of these ideas at different stages over the years:

Bethany Burum, Christine Looser, Gus Cooney, Kyle Dillon, Adam Mastroianni, Thalia Wheatley, Beau Sievers, David Amodio, Timothy Wilson, Bria Long, Mike Yeomans, Paul Seli, Hauke Hillebrandt, Fiery Cushman, Spencer Lynn, Scott Olesen, Annemarie Kocab, Alek Chakroff.

To the many people who encouraged and paved the way for me to enter graduate school to begin with: Penny Visser, George Wu, Eugene Caruso, Jane Risen, Nick Epley, Travis Carter, Nadav Klein, Carey Morewedge, Asia Eaton, Anne Henly, and Matt Patton. Most of all, to Liz Majka, without whom I would barely know what psychology was, let alone study it for a living.

Our department is housed in William James Hall, and is full of staff people who do their essential jobs so well that we all too often forget to appreciate them. They make all of our work possible. Thank you to Celia Raia, Mark Gerstel, Susan Kany, Joan Smeltzer, Regina Laskowski, Wendy Erselius, Katie Powers, Andrea Lynch, Kathy Kaufman, Mark Cellucci, Eudeen Green, Bill Santoro, Cindy Fiore, Jimmy, Anastasia, and Saintli.

To the not-so-small army of research assistants who did the painstaking work of running these experiments, and gave their sweat, tears, and way too many brain cells to staring at blue dots all day: Samantha Acker, Alexa Altchek, Mohin Banker, Sophie Carroll, Petrina Chan, Rachel Chmielinksi, Ashley Collinsworth, Joanne Crandall, Kat da Silva, Claire Dickson, Irene Droney, Cailey Fitzgerald, Shannon Ganley, Ashley Gong, Molly Graether, James Green, Hannah Harris, Lauren Harris, Uriel Heller, Tanner Hicks, Sarah Hoffman, Dallas Hogan, Vesper Hu, Emily Kemp, Benny Kollek, Rachel Lisner, Zoe Lu, Brenna Martinez, Tina Murphy, Aileen Navarrete, Debbie Park, Derek Peng, Michael Powell, Connor Richardson, Margo Sanders, Chace Shaw, Claire Shi, Jocelyn Skoler, Gemma Stern, Robin Stramp, Alexandro

Strauss, Laurel Symes, Harriet Tieh, Mikaela Thompson, Angel Wang, Iris Wang, Christie Wu, Xin Zeng, Stephanie Zatwarnicki, Yaojia Zheng, and Samuel Zwickel.

To my parents, who are the reason that I'm here, and who have worked hard their entire lives so that I could build a life of my own doing something that I love. I hope this makes up for that time in middle school when I wrote the class newsletter myself, and then didn't put my name on it. My name is definitely on this one.

To my wife, Tanya, and our three children, Noah, Sasha, and Maya, who I love more than anyone in the whole world. Noah was born as I started writing the first draft of what would become this dissertation, and Sasha and Maya were born as I started writing the final draft. I will never have better coauthors in my life.

INTRODUCTION

"Consistency is contrary to nature, contrary to life. The only completely consistent people are dead."

Aldous Huxley

"Men become accustomed to poison by degrees."

Victor Hugo

The deformation of a solid under load is known as *creep*. But in the last few years, that term has crept beyond material science and has come to describe almost any kind of unintended expansion of a boundary. Software developers worry about feature creep (the unintended expansion of a product's function over time), project managers worry about scope creep (the unintended expansion of a team's mandate over time), and military commanders worry about mission creep (the unintended expansion of a campaign's objectives over time). As it turns out, abstract concepts creep too. For example, in 1960, Webster's dictionary defined *aggression* as "an unprovoked attack or invasion," but today that concept can include behaviors such as making insufficient eye contact or asking people where they are from (Lilienfeld, 2017). Many other concepts, such as *abuse, bullying, mental disorder, trauma, addiction*, and *prejudice* have expanded of late as well (Haslam, 2016). Some see these expansions as an unwelcome sign of political correctness, while others see them as a welcome increase in social sensitivity. I take no position on whether these expansions are good or bad, but rather, seek to understand what makes them happen. Why do concepts creep?

I suggest that concepts naturally expand when instances of the concept become less prevalent—a phenomenon I call *prevalence-induced concept change.* Psychologists have long

known that stimuli are judged in the context of the relevant stimuli that precede them in time or surround them in space (for a review, see Schwartz, Hsu, & Dayan, 2007), and it is no surprise that the perceived aggressiveness of a particular behavior depends on the aggressiveness of the other behaviors the observer has seen. When instances of a concept become less prevalent—for example, when unprovoked attacks and invasions decline—the context in which new instances are judged changes as well. When most behaviors are less aggressive than they once were, some behaviors will naturally seem more aggressive than they once did, which may lead observers to mistakenly conclude that the "prevalence of aggression" has not decreased. Rather than disappearing when their instances do, concepts such as aggression may survive by expanding to include instances that they previously excluded.

Concepts and signal detection

Using concepts to identify and group stimuli is a particular case of a broader problem that organisms face: how to distinguish signals from noise. In the language of Signal Detection Theory (Tanner Jr & Swets, 1954), a signal is anything of interest to the observer, and noise is everything else. Signals can be desirable, like food, or dangerous, like predators. The observer uses sensory information to determine whether a particular stimulus is a signal or noise. A shopper can use color and softness to determine if a banana is ripe, and a musician can use pitch to figure out if a note played on the violin is flat or sharp. Sometimes a constellation of multiple cues is needed to identify a signal. A glass of milk might not be spoiled just because it smells slightly funny, but upon seeing the sluggish way it moves when the glass is tilted, one can pour it down the drain with confidence. Psychophysicists often collapse cues into a single measure of *intensity,* which represents all of the available information about a stimulus that the observer can use to identify it (Wickens, 2001). Stimuli with high intensity (yellow bananas and smelly milk)

would be labeled as signals, while those with low intensity (green bananas and fresh milk) would be labeled as noise.

The amount of intensity at which an observer starts identifying stimuli as signals rather than noise is called the observer's *criterion*. If two farmers are trying to find rotten strawberries, and farmer A has a liberal criterion while farmer B has a strict criterion, then farmer A will throw away more strawberries. Why? Because farmer A is willing to call slightly bruised strawberries rotten, while farmer B's criterion dictates that only the most black, bruised fruit will be discarded. In the absence of an explicit, external standard about which strategy is better, neither is wrong. Individual observers adopt a higher or lower criterion in order to satisfy different preferences for their search tasks (Macmillan & Creelman, 1991).

Terminology

In the present text, I will use *concept* to refer to any semantic category that an observer uses to identify and group similar stimuli. I will use *intensity* to refer to the actual amount of evidence that a particular belongs to concept A rather than concept B, and *threshold* or *decision threshold* to refer to the amount of intensity required before the observer identifies a stimulus as belong to concept B rather than concept A.[1] *Prevalence* will refer to the proportion of stimuli in the local environment from concept A, as opposed to concept B.

For example, imagine an observer whose task is to view colored dots and decide whether each dot is one particular color (blue) or another (purple). Using the terminology above, the colors "blue" and "purple" are concepts A and B, respectively. The color of a given dot is its

[1] Because the term *criterion* usually refers to one of several specific parameters calculated in Signal Detection Theory research, I will use the term *threshold* instead, since I do not use those parameters here.

intensity. If the observer only classifies dots as blue when, for example, they only exceed a certain blue value in RGB coordinates, that value is the observer's decision threshold.

In this example, what determines the prevalence of blue dots? Unlike a psychophysical task with objective stimuli (such line orientation, where "above 90 degrees" and "below 90 degrees" are discrete classes of stimuli that do not overlap), color is a continuous spectrum – there is no objective definition of where purple stops and blue begins. Instead, prevalence can be arbitrarily defined by the experimenter. Unless otherwise noted, the examples and studies described in the present work will set the threshold between two concepts at the midpoint of intensity between them, on whatever scale of intensity is being used.

When concepts should and should not change

Should decision thresholds be influenced by the prevalence of instances? For one class of decisions, the answer is yes. An emergency room physician may normally identify patients with gunshot wounds as "needing immediate attention" and patients with broken fingers or sore throats as "not needing immediate attention." But if on a particular day there happen to be no patients with gunshot wounds in the emergency room, it makes sense for the doctor to relax her threshold and start identifying patients with broken fingers as "needing immediate attention." It would be rather odd for the doctor to maintain her previous threshold and force patients with broken fingers to sit around and wait even when she has no patients with more serious injuries.

Decisions about who needs immediate care should be *prevalence-dependent*. But another class of decisions should be *prevalence-independent*. A doctor may normally identify patients with gunshot wounds and broken arms as "needing pain medication" and patients with sore throats as "not needing pain medication." But if on a particular day there happen to be no patients with gunshot wounds in the emergency room, it would be rather odd for the doctor to

relax her threshold and start identifying patients with sore throats as "needing pain medication." The threshold for administering pain medication should be held constant over time. There may be days when many patients require pain medication and there may be days when few require it, but it would be bad practice for a doctor to administer pain medication to a fixed number of patients every day, no matter what their conditions. Prevalence-independent decisions require that people hold their thresholds constant over time, and I suspect that this is something that people find difficult to do.[2]

BACKGROUND

What does previous work already tell us about the relationship between concept size and prevalence, and why might I expect concepts to change based on the prevalence of their instances? In this section, I will review conceptual and empirical work that lends support to the proposed phenomenon, as well as relevant findings that would predict otherwise.

Conceptual thresholds as defined here are similar to the *criterion* in Signal Detection Theory. In a scenario where an observer is trying to decide if a stimulus is an instance of concept B rather than concept A, their threshold of intensity can be thought of as a bias towards one concept and against the other. Given this similarity, it would be reasonable to ask what research using Signal Detection Theory says about the influence of signal prevalence on the observer's criterion. Two notable phenomena make direct predictions about this relationship.

The Vigilance Decrement

[2] Prevalence-independent decisions may be unique to humans. Other animals generally need not worry about consistency over time in their decisions, as long as they are maximizing their fitness, and accruing resources safely and efficiently. What an animal considers "edible" or a "dangerous predator" should and does change based on how scarce those things are in the environment (Trimmer, Ehlman, McNamara, & Sih, 2017).

The *vigilance decrement* is a term for the well-known finding that observers monitoring for rare signals suffer a decrease in detection accuracy as time goes on (Mackworth, 1948). In other words, the longer an observer is on the lookout for something rare, the more likely they are to miss it when it finally does appear. Many popular explanations for the vigilance decrement invoke fatigue or depleted resources (e.g. Warm, Parasuraman, & Matthews, 2008), while other researchers argue that fatigue causes an apparent decrease in the observer's sensitivity, or ability to distinguish signals from noise, (e.g. Matthews & Davies, 2001; Temple et al., 2000). Recently, Thomson and colleagues (2015) have posited that the vigilance decrement is actually due to the observer's criterion growing more conservative over time. According to this account, observers grow progressively more inattentive as time goes on, requiring a more intense stimulus in order to capture attention.

A criterion-based explanation of the vigilance decrement seems at odds with my predictions. I predict that participants instead become more likely to identify stimuli as signals when their prevalence decreases, which would result in more instances of incorrectly classifying stimuli as signals that they previously classified as noise. Additionally, an expansion of conceptual boundaries would essentially be a liberal criterion shift, which is the opposite of the vigilance decrement mechanism that Thomson and colleagues propose.

The Low Prevalence Effect

The low-prevalence effect (Wolfe et al., 2007) describes a failure in human observers' ability to correctly identify rare targets in visual search. However, while the vigilance decrement focuses on the harmful effects of prolonged observation, the low-prevalence effect suggests that people have trouble accurately detecting rare signals, for the very reason that they are rare. As

the researchers have put it, “if you don’t find it often, you often don’t find it” (Evans, Birdwell, & Wolfe, 2013).

Unlike vigilance tasks, which involve long periods of waiting for a signal to appear onscreen, LPE paradigms use complex visual search tasks, in which each trial is an array of visual objects presented at once, with a signal either present or absent among many distractor items. Examples include searching for the letter L among a large group of Ts (Godwin, Menneer, Riggs, Cave, & Donnelly, 2014), looking for a butterfly in an array of other animals (Hout, Walenchok, Goldinger, & Wolfe, 2015), or looking for weapons in an x-ray of a packed piece of luggage (Wolfe, Brunelli, Rubinstein, & Horowitz, 2013). As Wolfe and colleagues (2007) point out, while the observer in vigilance tasks has to be on guard for a signal that suddenly appears and then disappears (and might be missed due to inattention), an observer in an LPE task has as much time as they need to look for the signal in each array, and to decide when they have looked long enough to deem a trial “signal absent.”

What causes the low-prevalence effect? Wolfe and van Wert (2010) argued that it is primarily due to a shift in the observer’s criterion, which can be described by a drift diffusion model (Ratcliff, Smith, Brown, & McKoon, 2016). At low prevalence, their model predicts, observers adopt an earlier “quitting time” and are not willing to spend as much time searching for signals. This finding is supported in a variety of different types of search tasks (e.g. Peltier & Becker, 2016). However, given the complex nature of real-world visual search, it is unsurprising that the LPE may have multiple causes. Other contributing factors may include motor errors (Mitroff & Biggs, 2013), lack of confidence in perception during low prevalence conditions (Schwark, Sandry, & Dolgov, 2013; Schwark, Sandry, MacDonald, & Dolgov, 2012), and basic, low level perceptual failures (Hout et al., 2015).

The LPE would seem to be at odds with my predictions about how prevalence shifts observer detection thresholds. A hallmark of many LPE studies is that observers adopt a more conservative criterion at low prevalence (Schwark et al., 2012; Wolfe et al., 2007, 2013; Wolfe & Van Wert, 2010). I instead predict that observers would relax their decision threshold when signals of interest become rare.

Visual aftereffects and sensory adaptation

The brain can adapt perception of the world based on recent input (see Helson, 1964; Webster, 2015). For example, humans adjust perceptions of luminance to preserve relative intensities, rather than trying to perceive physical luminance levels with perfect fidelity (Bartlett, 1965). Adaptation is thought to be an evolutionarily advantageous feature of sensory systems, since the statistical regularities of the natural world mean that focusing on differences between recent inputs is an efficient way to encode sensory information (Schwartz et al., 2007).

Many studies of adaptation have focused on the case of visual aftereffects, or distortions in perception produced by viewing a stimulus for a long time (for a review, see Thompson & Burr, 2009). Prolonged exposure to photographs of artificially broadened faces causes normal faces to seem artificially narrow or pinched (Leopold, Rhodes, Muller, & Jeffery, 2005; Rhodes, Jeffery, Watson, Clifford, & Nakayama, 2003). Similar "repulsive" aftereffects have been documented for colors (M. Webster, 1996), motion (Anstis, Verstraten, & Mather, 1998; Mather, Pavan, Campana, & Casco, 2008) and spatial orientation (Paradiso, Shimojo, & Nakayama, 1989). However, there is also evidence of adaptation occurring in the opposite direction in domains such as numerosity perception (Fornaciai & Park, 2018).

I would argue that the body of evidence for repulsive aftereffects supports my predictions about the relationship between prevalence and decision thresholds, at least in visual perception.

A repulsive aftereffect is essentially a perceptual bias away from the mean of recent stimuli. If reducing the prevalence of very blue dots causes observers to call dots in the middle of the spectrum blue, they are in essence biasing their classifications away from the historical mean stimulus intensity (very blue).

Beyond the extensive literature on facial adaptation and aftereffects, recent work arguing that adaptation also takes place for decision processes more generally (Cheadle et al., 2014) suggests that an adaptation account of prevalence-induced concept change could extend non-perceptual domains, such as moral judgment. While both traditional models of signal detection and basic Bayesian models suggest that rare events should be biased against since they are unexpected (see Summerfield & de Lange, 2014), recent Bayesian models of adaptation (Clifford et al., 2007) and work on related neural computations such normalization (e.g. Carandini & Heeger, 2012; Louie et al., 2015) could be promising candidates for models that accurately describe my predictions.

Contextual effects in judgment

There is an extensive body of work on the relative nature of human judgments, which are often made in comparison to recent or salient stimuli (e.g. Kahneman, Miller, Griffin, Mcpherson, & Read, 1986; Mellers & Birnbaum, 1983; Tversky & Simonson, 1993). Anchoring (Epley & Gilovich, 2006) is the well-known tendency for numerical estimates to be biased towards recently encountered numbers that are available in memory, even when they are irrelevant to the current estimate. Work on reference points in valuation broadly (Tversky & Kahneman, 1991), as well as their use in specific domains like risk-taking (Schneider, 2016) and goal pursuit (Heath, Larrick, & Wu, 1999) show that the subjective magnitude of a value, or

achievability of some target level of performance, can feel larger or smaller based on salient values to which they are compared.

Contrast effects are another extensively studied contextual influence on judgment. In situations where a decision-maker is evaluating many stimuli sequentially, valuations of the current stimulus often depend on the values of recently viewed stimuli. As the name suggests, contrast effects involve a repulsive bias, pushing the perceived value of the current stimulus away from recently experienced values. For example, Bhargava and Fisman (2014) found that ratings of current speed-dating partners were biased away from the likability ratings of past partners. Similar effects have been documented with populations as diverse as Olympic athletes (Damisch, Mussweiler, & Plessner, 2006), asylum judges, loan officers, and professional baseball umpires (Chen, Moskowitz, & Shue, 2016). In most of these lines of research, the contrastive bias is only significantly predicted by the most recent one or two stimuli in memory.

Context can mean many things depending on the question of interest. Often researchers studying such effects make a useful distinction between *spatial context* and *temporal context* (see Louie & Glimcher, 2012). Spatial context involves stimuli that are present at the time of judgment of the current stimulus. Temporal context refers to the prior history of context in memory that impacts the current judgment. The impact of the local prevalence in sequential judgment tasks would be considered an effect of temporal context.

Expectation and Bayesian Inference

Many researchers in the cognitive sciences have in recent years begun examining which mental operations in the brain can be described with Bayesian models of cognition (Griffiths, Kemp, & Tenenbaum, 2007). In these models, the brain tries to solve a problem with an approximation of Bayesian inference – by using prior knowledge from past events in order to

inform the predictions of future events. Indeed, many classic cases of Bayesian inference use prevalence information in order to illustrate how prior knowledge can usefully inform 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 considering base rates of the color of cars in the city.

Related lines of research in perceptual decision making have focused on the roles of priming effects (Maljkovic & Nakayama, 1994; Wiggs & Martin, 1998) and expectation (Summerfield, Egner, Greene, Koechlin, & Hirsch, 2015). In both phenomena, perceptual systems use recent experience or stimuli to more efficiently and accurately interpret incoming sensory information. Bayesian or quasi-Bayesian processes may be involved in both priming (Schooler, Shiffrin, & Raaijmakers, 2001) as well as in expectation effects (Summerfield & de Lange, 2014).

Are existing Bayesian models of perception compatible with my predictions about how prevalence will change concepts in human judgment? It depends on the type of inference being made. In a simple model of Bayesian probability estimation applied to the color example I described earlier, imagine that an observer in a color identification task is trying to guess the probability that a dot is blue. In this scenario, the prior used to update the posterior probability that the dot is blue is the base rate of blue dots. When blue dots are rare, the prior goes down, and so does the posterior probability. In other words, this kind of Bayesian account would suggest that as blue dots become more rare, the observer should be biased against finding more blue dots – exactly the opposite of my predictions.

In a second kind of Bayesian inference, 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 perception of that average to recent history. Such models (e.g. Wei & Stocker, 2015) would predict, as I do, that observers are more likely to call dots blue when bluer dots become rare.

Concept creep

Is there evidence of concepts expanding or contracting on a societal level? Haslam (2016) argued with a combination of observational and archival data that many concepts important to psychology, including *bullying, mental disorder,* and *trauma* have expanded their borders over the past several decades, even while the prevalence of instances of those concepts have either remained stable or declined over the same time period. Haslam does not make a claim as to a singular mechanism for this phenomenon, perhaps because he focuses on the type of conceptual change he can observe most readily in the world – semantic change. As an example, Haslam argues that the set of behaviors commonly referred to as *aggressive* has broadened over the past several decades. He remains agnostic as to whether this change in the semantic boundaries of the term *aggression* is accompanied by an actual change in the underlying behavior, or in perception of the behavior. He discusses several possible explanations for the phenomenon, but all of them are conscious, cultural or societal forces, rather than lower-level cognitive processes.

I predict that prevalence-induced concept change happens in a specific direction: concepts should expand when instances grow rare, and conversely, contract when instances become common. Haslam argues for six examples of concept creep he has observed in the real world – abuse, bullying, trauma, mental disorder, addiction, and prejudice – and in none of these categories does he argue (or does the archival data he cites suggest) that any of these concepts have grown more prevalent over time, which would contradict my prediction. In other words,

while prevalence-induced concept change is unlikely to be the sole explanation for the real-world examples of concept creep that Haslam cites, it is compatible with all of them.

Overview of the present research

The studies included here are designed to present evidence for prevalence-induced concept change across several domains – color perception (Studies 1 and 2), facial threat perception (Study 7), and ethical judgments (Study 8). I also attempt a series of interventions to remove or mitigate the influence of prevalence changes on concept change in the color domain, largely without success (Studies 3-6, 9). I conclude by presenting several candidate mechanisms for the process that drives prevalence-induced concept change, and introducing a computational model which suggests that the most likely cognitive mechanism for the phenomenon is Range-Frequency Theory (Parducci, 1965).

STUDY 1: DECREASING PREVALENCE OF COLORS

Overview

Do observers shift their conceptual boundaries when instances of a concept become more rare or common over time? Study 1 was designed to answer this basic question. Participants viewed a series of dots on a computer screen and were asked to identify each dot as either blue or not blue. After many trials, the prevalence of blue dots decreased for some participants. This and all subsequent studies were approved by the Harvard University Committee on the Use of Human Subjects.

Methods

Sample. Participants were 22 students at Harvard University (6 males, 16 females, M_{age} = 22.5 years, SD = 1.9 years) who received either money or course credit in exchange for their

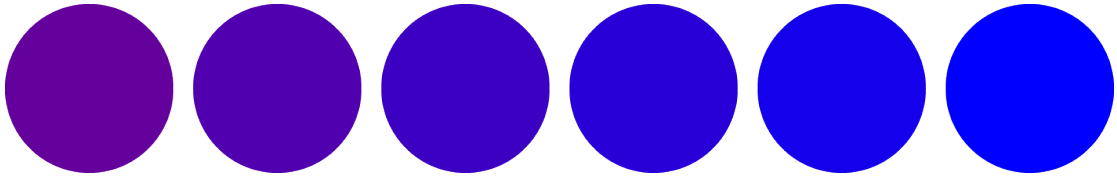
participation. One female participant experienced a minor medical problem during the study and her data were excluded, leaving 21 participants in the data set. In this and all subsequent studies: (a) I set a minimum sample size based on previous research that had used similar methods and stimuli, (b) once I reached the minimum sample size, I continued to recruit participants through the end of the academic term, (c) I did not analyze my data until all participants had been recruited, (d) all manipulations, measures, and data exclusions are reported, and (e) data exclusions had no impact on the significance of the results.

*Procedure***.** Upon arrival at the laboratory, participants were escorted to a room equipped with a computer display and keyboard, and they remained there for the duration of the study. 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 that were respectively labeled "blue" and "not blue."

On each trial, a colored dot appeared on a solid gray background. The color of the dot varied across trials from very purple (60% blue, RGB 100-0-155) to very blue (99.6% blue, RGB 1-0-254). 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 1000 trials divided into 20 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 1000 test trials. To help participants remain attentive, I allowed them to take a break every 50 trials.

I created two conditions by dividing the color spectrum into two halves that I will refer to as the "purple spectrum" (RGB 100-0-155 through RGB 51-0-204) and the "blue spectrum" (RGB 50-0-205 through RGB 1-0-254), as shown in Figure 1. 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. 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 1-1000 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 351-1000. After completing the identification task, participants completed a questionnaire asking some basic demographics and their impressions of the task (see Appendix II).

FIGURE 1: Examples of Dots Used in Color Identification Task



The color spectrum comprised 100 dots ranging from approximately RGB 100-0-155 (very purple) to RGB 0-0-255 (very blue) and this figure shows (from left to right) the 1st, 20th, 40th, 60th, 80th, and 100th dots. The three dots on the left are from the purple spectrum and the three dots on the right are from the blue spectrum.

Results

The tasks that participants performed in this and all subsequent studies performed may be thought of as signal detection tasks. However, traditional signal detection tasks present participants with stimuli that can be objectively classified as either signal or noise, and the data are typically analyzed by using the number of correct and incorrect responses to calculate *d′* (sensitivity) and *c* (response threshold) for each participant. Because there are no "objectively correct" answers to questions such as "Is this dot blue?" or "Is this face threatening?" or "Is this proposal acceptable?" it is not possible to calculate these traditional parameters for my data. My alternative analytic approach is described below.

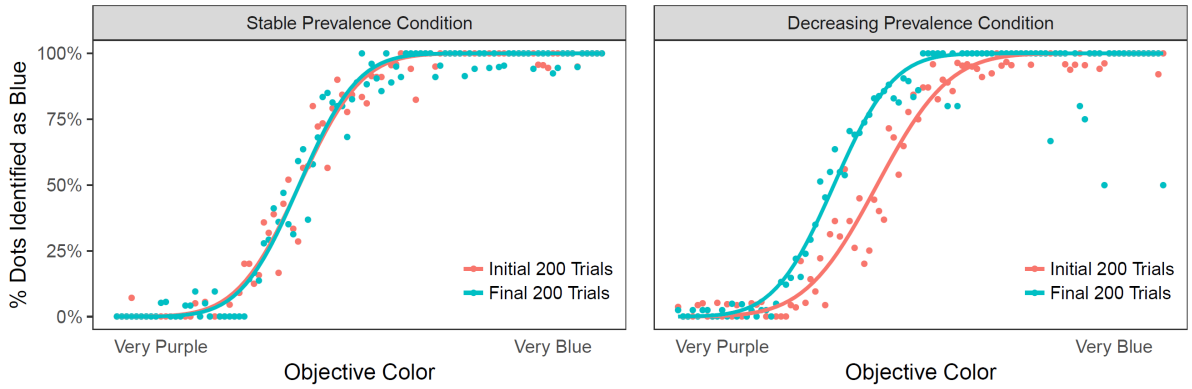
Did the decrease in the prevalence of blue dots cause participants' concepts of *blue* to expand? To find out, I fit a binomial generalized linear mixed model to my data in R (R Core Team, 2017) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). The dependent variable was the participant's *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 *objective color* (which ranged from 0% blue to 99.6% blue) and (b) the *trial number* (which ranged from 1 to 1000). I included condition, trial number, and objective color (and all interactions between them) as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 494.59$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 127.66$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between

condition, trial number, and objective threateningness significantly improved model fit, $\chi^2(1) = 48.34$, $p < 0.001$.

The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications, $b = 12.50$, $SE = 1.75$, $z = 7.14$, $p < 0.001$, *95% CI* [8.85, 16.09], $R^2_{GLMM(c)} = 0.88$.[3] (All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples). Figure 2 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' concepts of *blue* expanded to include dots that it had previously excluded.

[3] In all models presented here, $R^2_{GLMM(c)}$ (the conditional pseudo-r-squared) is calculated as described by Nakagawa and colleagues (2012) using the piecewiseSEM package (Lefcheck, 2016).

FIGURE 2: Results for Study 1



The x-axes show the dot's objective color and the y-axes show the percentage of trials on which participants identified that dot as blue.

Conclusion

Participants who experienced a decrease in the prevalence of blue dots 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' concepts of *blue* expanded to include dots that it had previously excluded. This study did not address whether this kind of concept change was limited to prevalence decreases, or if it could occur with any kind of prevalence change. Study 2 was designed to address this question.

STUDY 2: INCREASING PREVALENCE OF COLORS

Overview

Study 1 demonstrated a tendency for observers to expand their concepts of a search target when instances of it grow rarer over time. Is this phenomenon specific to prevalence decreases, or is it a more general response to any prevalence change? To explore this question, I ran a study

similar to Study 1, but replaced the decreasing prevalence condition with a condition where the prevalence instead increased over time. Participants viewed a series of dots on a computer screen and were asked to identify each dot as either blue or not blue. After many trials, the prevalence of blue dots increased for some participants.

Methods

Sample. Participants were 23 students at Harvard University (11 males, 12 females, M_{age} = 22.1 years, SD = 2.5 years) who received course credit in exchange for their participation. One female participant did not follow the experimenter's instructions during the study and her data were excluded, leaving 22 participants in the data set.

Procedure. The method for Study 2 was virtually identical to the method for Study 1 except that I replaced the decreasing condition with an *increasing condition*. The signal prevalence in the increasing condition was 6% on trials 1-200; 16% on trials 201-250; 28% on trials 251-300; 40% on trials 301-350; and 50% on trials 351-1000. After completing the identification task, participants completed a questionnaire asking some basic demographics and their impressions of the task (see Appendix II).

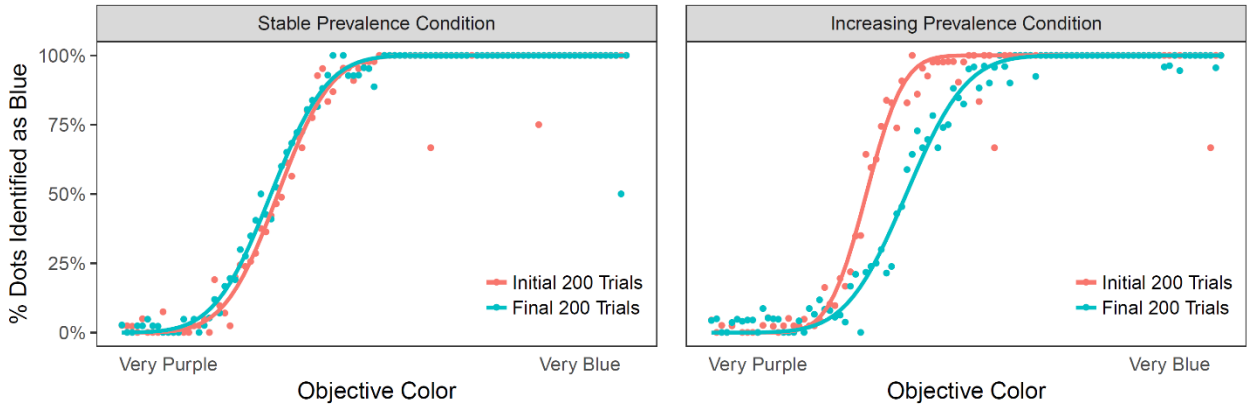
Results

Did the increase in the prevalence of blue dots cause participants' concepts of *blue* to contract (rather than to expand)? To find out, I fit a binomial generalized linear mixed model to my data in R using the lme4 package. 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 increasing). The independent within-participants variables were (a) the dot's RGB value or what I will call its *objective color* (which ranged from 0% blue to 100% blue) and (b) the *trial number* (which ranged from 1 to 1000). I included condition, trial number, and

objective color (and all interactions between them) as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. The inclusion of random slopes significantly improved model fit relative to the baseline model, $\chi^2(2) = 49.57$, $p < 0.001$, as did the inclusion of random intercepts, $\chi^2(2) = 386.15$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(1) = 15.12$, $p < 0.001$.

The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications, $b = -8.13$, $SE = 1.40$, $z = -5.83$, *95% CI* [-12.31, -4.05]], $R^2_{GLMM(c)} = 0.89$. (All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples). Figure 3 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-200) and the final trials (800-1000). The positive slope of all curves indicates that in both conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But the two panels differ in an important way. 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 a final trial as when it appeared on an initial trial. But the two curves in the right panel are offset in the middle, indicating that participants in the increasing condition were less likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial. In short, when blue dots became more prevalent, participants identified as not blue some dots that they had earlier identified as blue.

FIGURE 3: Results for Study 2



The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

Conclusion

Participants who experienced an increase in the prevalence of blue dots were less 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 increased, participants' concepts of *blue* contracted to exclude dots that it had previously included.

The results of studies 1 and 2 suggest that participants do not use information about their original decision threshold of a target to keep that threshold from moving after a prevalence change, at least in the domain of color perception. One potential reason for this is that participants in Studies 1 and 2 may have been surprised by the prevalence deviating from 50%, even when they were generally warned that the prevalence could change at any time. To further investigate whether this was a concern, and whether I could observe any conscious control over target expansion, in Study 3 I tried to explicitly warn participants about how the prevalence

would change. I predicted that warning participants about a prevalence decrease beforehand would prevent them from expanding their conceptual boundaries.

STUDY 3: FOREWARNING OF A PREVALENCE DECREASE

Overview

In Study 3, I replicated the procedure for Study 1, except that instead of telling participants in the decreasing condition that the prevalence of blue dots *might change* over trials, I told them that the prevalence of blue dots *would decrease* over trials.

Methods

Sample. Participants were 43 students at Harvard University (10 males, 31 females, M_{age} = 20.4 years, SD = 2.1 years) who received either money or course credit in exchange for their participation. Two female participants who were given incorrect study materials were excluded, as was one male participant who disregarded experimental instructions and one male participant who reported being colorblind. This left 39 participants in the data set.

Procedure. The method for Study 3 was identical to the method for Study 1 except that before the study began, participants were explicitly told what would happen to the prevalence of blue dots during the study. Participants in the decreasing condition were told: "As the study goes on, blue dots are going to become less common. In other words, you will see fewer of them over time." Participants in the stable condition were told: "As the study goes on, blue dots are not going to become more or less common. In other words, you will see the same amount of them over time." After completing the identification task, participants completed a questionnaire asking some basic demographics and their impressions of the task (see Appendix II).

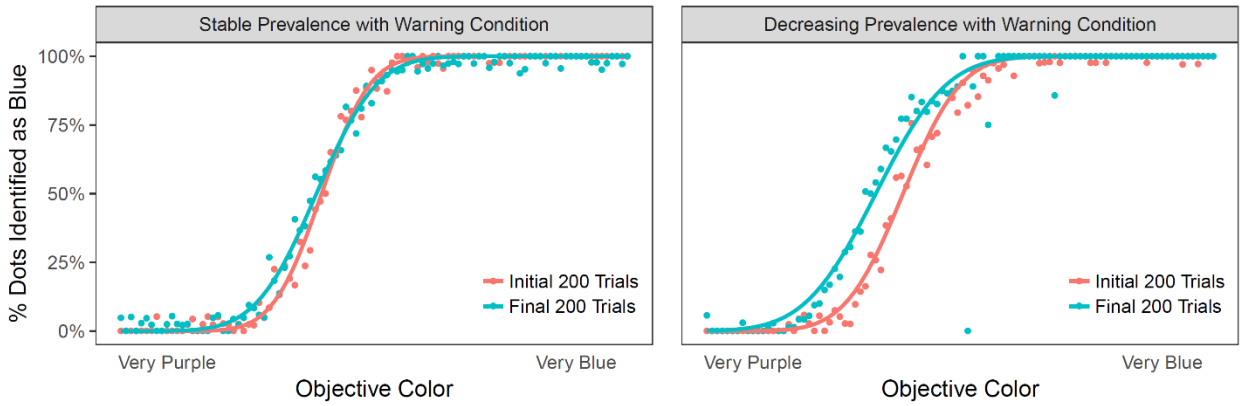
Results

Did the decrease in the prevalence of blue dots cause participants' concepts of *blue* to expand even when they were explicitly told that the prevalence of blue dots would decrease? To find out, I fit a binomial generalized linear mixed model to my data in R using the lme4 package. 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 *objective color* (which ranged from 0% blue to 99.6% blue) and (b) the *trial number* (which ranged from 1 to 1000). I included condition, trial number, and objective color (and all interactions between them) as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 692.36$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 229.86$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(1) = 117.91$, $p < 0.001$. The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications, $b = 21.74$, $SE = 1.55$, $z = 14.00$, *95% CI* [17.83, 25.77], $R^2_{GLMM(c)} = 0.93$. All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples.

Figure 4 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-200) and the final trials (800-1000). The positive slope of all curves indicates that in both conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But the two panels differ in an important way. 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 a final trial as when it appeared on an initial trial. But the two curves in the right panel are offset in the middle, indicating that participants in the decreasing condition were more likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial. In short, when blue dots became less prevalent, participants identified as blue some dots that they had earlier identified as not blue, and they did this even when they were explicitly warned about the decrease in prevalence.

FIGURE 4: Results for Study 3



The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

Conclusion

Participants who experienced a decrease in the prevalence of blue dots 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, even when they were forewarned about the prevalence change.

These results suggest that mere knowledge of an impending prevalence change does not prevent observers from shifting their decision thresholds for calling colors blue when they occur. However, it is unclear whether observers were unable to change their behavior, or were not sufficiently motivated to do so. Study 4 was designed to better motivate consistency throughout the color task.

STUDY 4: APPEAL AND FINANCIAL INCENTIVES FOR CONSISTENCY

Overview

In Study 4, I replicated the procedure for Study 1, except that this time a third of the participants in the decreasing condition were *explicitly instructed* not to change their identifications of dots over the course of the study ("Do your best to respond the same way if you see it again later in the study"), and another third were given the same explicit instruction and also offered a *monetary reward* for following it ("We will be awarding a bonus of $10 to the five most consistent participants in this study").

Methods

Sample. Participants were 92 students at Harvard University (34 males, 57 females, M_{age} = 18.4 years, SD = 2.1 years) who received course credit in exchange for their participation. One female participant who was interrupted during the experimental session was excluded, leaving 91 participants in the data set.

Procedure. The method for Study 4 was virtually identical to the method for Study 1 except for two things. First, I added two new conditions. Whereas participants in the *stable condition* and the *decreasing* condition were given the same instructions as they were given in Study 1, participants in the new conditions were given different instructions. Specifically,

participants in the new *decreasing+instruction* condition were told that once they had identified a dot as blue or not blue "you should do your best to respond the same way if you see it again later in the study." Participants in the new *decreasing+instruction+incentive* condition were told the same thing, and in addition, they were also told that "as an incentive, I will be awarding a bonus of $10 to the five most consistent participants in this study." The second change to the method of Study 1 is that I reduced the number of trials from 1000 to 800. As such, the signal prevalence in the stable condition was 50% on trials 1-800, and the signal prevalence in the decreasing condition, the decreasing+instruction condition, and the decreasing+instruction+incentive condition 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 (see Appendix II).

Results

Did the decrease in the prevalence of blue dots cause participants' concepts of *blue* to expand even when they were instructed, or instructed and incentivized, not to let this happen? To find out, I fit a binomial generalized linear mixed model to my data in R using the lme4 package. 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 *objective color* (which ranged from 0% blue to 100% blue) and (b) the *trial number* (which ranged from 1 to 800). I included condition, trial number, and objective color (and all interactions between them) as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b)

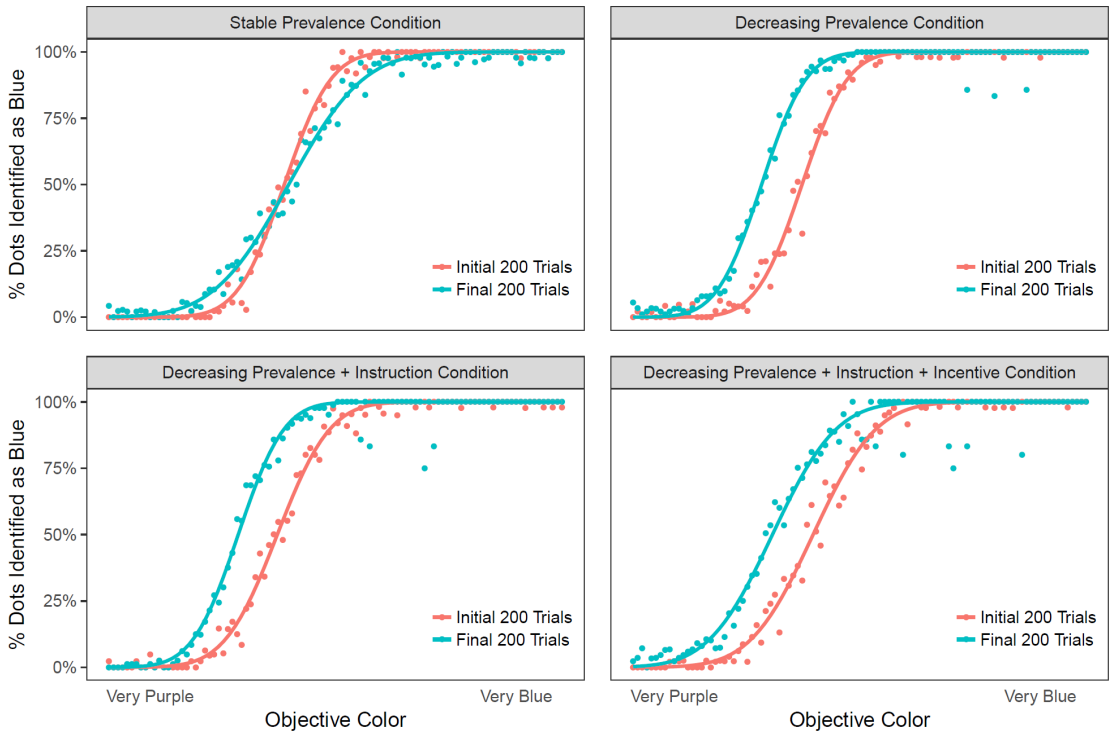
slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 1084.00$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 500.29$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(3) = 234.53$, $p < 0.001$.

The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications. Specifically, the stable prevalence condition differed significantly from the decreasing prevalence condition, $b = 21.98$, $SE = 0.67$, $z = 32.8$, $p < 0.001$, *95% CI* [18.44, 25.49], $R^2_{GLMM(c)} = 0.93$, the *decreasing* + instruction condition, $b = 27.84$, $SE = 1.48$, $z = 18.8$, $p < 0.001$, *95% CI* [23.72, 31.88], and the *decreasing* + *instruction* + *incentive* condition, $b = 15.34$, $SE = 1.29$, $z = 11.9$, $p < 0.001$, *95% CI* [12.13, 18.38]. The *decreasing* + *instruction* condition also differed significantly from the decreasing prevalence condition, $b = -5.86$, $SE = 0.71$, $z = -8.3$, $p < 0.001$, *95% CI* [-9.78, -1.91], as well as from the *decreasing* + *instruction* + *incentive* condition, $b = -12.50$, $SE = 1.09$, $z = -11.5$, $p < 0.001$, *95% CI* [-16.14, -8.85]. Finally, the *decreasing* + *instruction* + *incentive* condition differed significantly from the decreasing prevalence condition, $b = 6.64$, $SE = 0.67$, $z = 9.9$, $p < 0.001$, *95% CI* [3.56, 9.94]. (All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples, and all reported p-values are adjusted for multiple comparisons using the Holm correction).

Figure 5 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-200) and the final trials (600-800). The positive slope of all curves indicates that in all conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But the panels differ in an important way. The two

curves in the upper 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 a final trial as when it appeared on an initial trial. But in each of the other three panels, the two curves are offset in the middle, indicating that participants in the three decreasing conditions were more likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial. In short, when blue dots became less prevalent, participants identified as blue some dots that they had earlier identified as not blue, even when they had been instructed and incentivized not to let that happen.

FIGURE 5: Results for Study 4



The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

Conclusion

Participants who experienced a decrease in the prevalence of blue dots 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, even when they were exhorted to remain consistent in their identifications, and even when they were offered a financial incentive to do so.

Taken together, Studies 3 and 4 suggest that the impact of prevalence on concept change is not easily modulated by conscious effort. Participants in these studies were aware of prevalence changes or explicitly asked to not let their classifications change, but did so anyway. Isolating the particular features of the identification task and prevalence shifts that lead to conceptual change may reveal more about its underlying cause. Studies 5 and 6 were designed to address this topic.

STUDY 5: GRADUAL AND ABRUPT PREVALENCE DECREASES

Overview

Could the speed of a prevalence change affect how or whether observers shift their thresholds? Perhaps observers who experience a gradual prevalence change, like those in Studies 1-4, do not realize in real time that the prevalence is changing, and respond by shifting their thresholds in an attempt to maintain the previous rate of detection of a concept, even as instances of that concept become more rare. If this account is correct, then an abrupt prevalence change might not cause a threshold shift, since participants would notice the change in prevalence, and realize that a corresponding change in the rate of detection is appropriate.

In Study 5, I replicated the procedure for Study 1, except that in Study 5 I decreased the prevalence of blue dots *gradually* for some participants (as I did in the previous studies) and *abruptly* for others.

Procedure

Sample. Participants were 37 students at Harvard University (12 males, 25 females, M_{age} = 19.4 years, SD = 1.5 years) who received either money or course credit in exchange for their participation.

Procedure. The method for Study 5 was virtually identical to the method for Study 1 except for two things. First, I reduced the number of trials from 1000 to 800. Second, I added a new condition. For participants in the *stable condition,* the signal prevalence on trials 1-800 was 50%. This condition was the same as the stable prevalence condition in Study 1. For participants in the *gradually 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. This condition was the same as the decreasing condition in Study 1. For participants in the new *abruptly decreasing* condition, the signal prevalence was 50% on trials 1-200, and 6% on trials 201-800. After completing the identification task, participants completed a questionnaire asking some basic demographics and their impressions of the task (see Appendix II).

Results

Did the decrease in the prevalence of blue dots cause participants' concept of *blue* to expand even when the decrease occurred abruptly? To find out, I fit a binomial generalized linear mixed model to my data in R using the lme4 package. 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, gradually decreasing, or abruptly decreasing). The independent within-participants variables were (a) the dot's RGB value or what I will call its *objective color* (which ranged from 0% blue to 100% blue) and (b) the *trial number* (which ranged from 1 to 800). I included condition, trial number, and objective color (and all

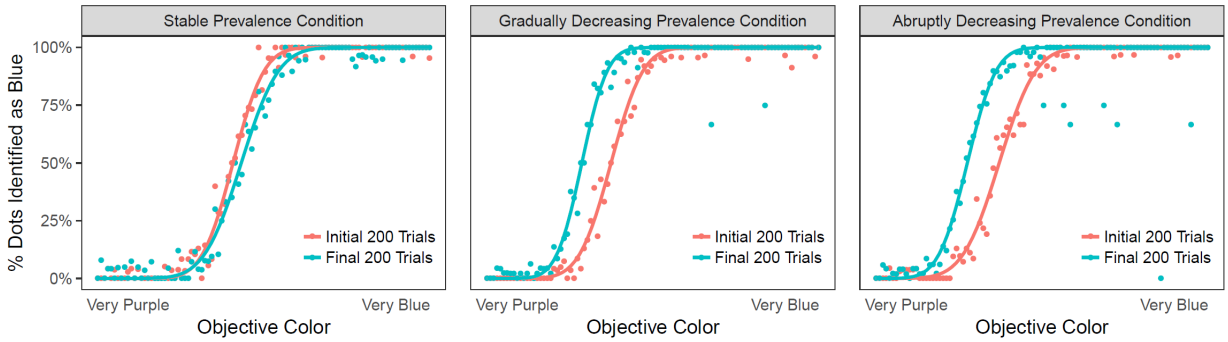
interactions between them) as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 234.49$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 48.32$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(1) = 72.52$, $p < 0.001$.

The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications. Specifically, the stable prevalence condition differed significantly from both the gradually decreasing prevalence condition, $b = 15.92$, $SE = 1.19$, $z = 13.3$, $p < 0.001$, *95% CI* [*11.63*, 20.36]], $R^2_{GLMM(c)} = 0.89$, as well as from the abruptly decreasing prevalence condition, $b = 15.26$, $SE = 0.56$, $z = 27.3$, $p < 0.001$, *95% CI* [11.04, 19.45]. However, the gradually and abruptly decreasing prevalence conditions did not differ significantly from one another, $b = 0.66$, $SE = 1.19$, $z = 0.6$, $p = 0.58$, *95% CI* [-3.96, 5.34]. (All reported 95% confidence intervals are the result of a bootstrapping procedure using 1000 bootstrap samples, and all reported p-values are adjusted for multiple comparisons using the Holm correction).

Figure 6 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-200) and the final trials (600-800). The positive slope of all curves indicates that in both conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But the two right panels differ in an important way. 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 a final trial as when it appeared on an initial trial. But the two curves in each of the two right panels are offset in the middle, indicating that participants in the two decreasing conditions were more likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial. In short, when blue dots became less prevalent, participants identified as blue some dots that they had earlier identified as not blue, and they did this even when the decrease in prevalence happened abruptly.

FIGURE 6: Results for Study 5



The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

Conclusion

Participants who experienced a decrease in the prevalence of blue dots 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, whether the decrease in prevalence was gradual or abrupt. This suggests that gradual speed of prevalence changes in Studies 1-4 were not responsible for this effect, and that

observers still shift their thresholds even when a change in prevalence is abrupt, and thus, very noticeable.

STUDY 6: VISUAL REFERENCE

Overview

Could participants who are experiencing a prevalence shift resist changing their decision thresholds if they had a reminder of their original thresholds? To investigate this, I presented participants in Study 6 with a visual reference – their own personal boundary between blue and purple. I predicted that explicitly showing participants their own boundary would prompt them to keep that it consistent over time, even if the prevalence of colors they saw changed.

Procedure

Sample. Participants were 70 students at Harvard University (24 males, 45 females, 1 gender specified, M_{age} = 19.5 years, SD = 1.5 years) who received course credit in exchange for their participation. One participant was excluded who failed to finish the study. This left 69 participants in the data set.

Procedure. In Study 6, I replicated Study 1 by varying the prevalence of blue dots across subjects, but also manipulated an additional between-subjects variable, the presence of absence of a visual reference onscreen, to produce a 2x2 design with four conditions. The first two conditions, the *stable + no reference* condition and the *decreasing + no reference* condition, respectively, replicated the two conditions of Study 1. The third condition, *stable + reference,* kept the prevalence of blue dots at 50% while adding a visual reference (see details below) onscreen on every trial. The fourth condition, *decreasing + reference,* slowly decreased the

prevalence of blue dots from 50% to 6% while also adding a visual reference onscreen with every trial.

Subjects evaluated 800 dots ranging from very blue to very purple. I manipulated prevalence between subjects, in the same manner as study 1, with a stable and a decreasing prevalence condition. However, after participants in either condition saw and evaluated 100 dots, their responses were fit to a cumulative normal function in real time in order to find their *point of subjective equality* (PSE) between blue and purple – the color that they found maximally ambiguous. This color served as a proxy for the participant's own threshold between blue and purple.

In the two trials with visual references, after the first 100 trials, participants began to see two dots on the screen. On the right-hand side of the screen was the target dot which they were to identify as being blue or not blue. On the left-hand side of the screen, another dot was present for all trials. Its color was the shade representing the subject's PSE, as calculated by a cumulative normal function fitted to the responses from the participant's first 100 trials. Additionally, I also changed the way I manipulated prevalence. In previous studies, prevalence was manipulated by showing more or fewer dots from above or below the arithmetic midpoint of my color spectrum (RGB 50-0-205). In this study, prevalence was instead varied around the subject's own PSE. Over time, participants in the decreasing prevalence condition saw fewer dots that were more blue than their PSE. Participants were informed after trial 100 that the color representing their definition of the blue would be present on the left-hand side of the screen for the remainder of the study. They were told that this color was the least blue color that they still called blue, or in other words, their personal definition of where the color blue began. After completing the

identification task, participants completed a questionnaire asking some basic demographics and their impressions of the task (see Appendix II).

Results

Did the presence of a visual reference cause participants to attenuate or eliminate their threshold shift in response to decreasing prevalence? To find out, I fit a binomial generalized linear mixed model to my data in R using the lme4 package. 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 + no reference, decreasing + no reference, stable + reference* or *decreasing + reference*). The independent within-participants variables were (a) the dot's RGB value or what I will call its *objective color* (which ranged from 0% blue to 100% blue) and (b) the *trial number* (which ranged from 1 to 800). I included condition, trial number, and objective color (and all interactions between them) as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. The inclusion of random slopes significantly improved model fit relative to the baseline model, $\chi^2(2) = 110$, $p < 0.001$, as did the inclusion of random intercepts, $\chi^2(2) = 1063$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(3) = 76.4$, $p < 0.001$.

The generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications, $R^2_{GLMM(c)} = 0.93$. Specifically, the *stable + reference* condition differed significantly from the *decreasing + reference* condition, $b = 15.06$, $SE = 0.61$, $z = 24.56$, $p < 0.001$, *95% CI* [13.86, 16.26], , the *stable + no reference*

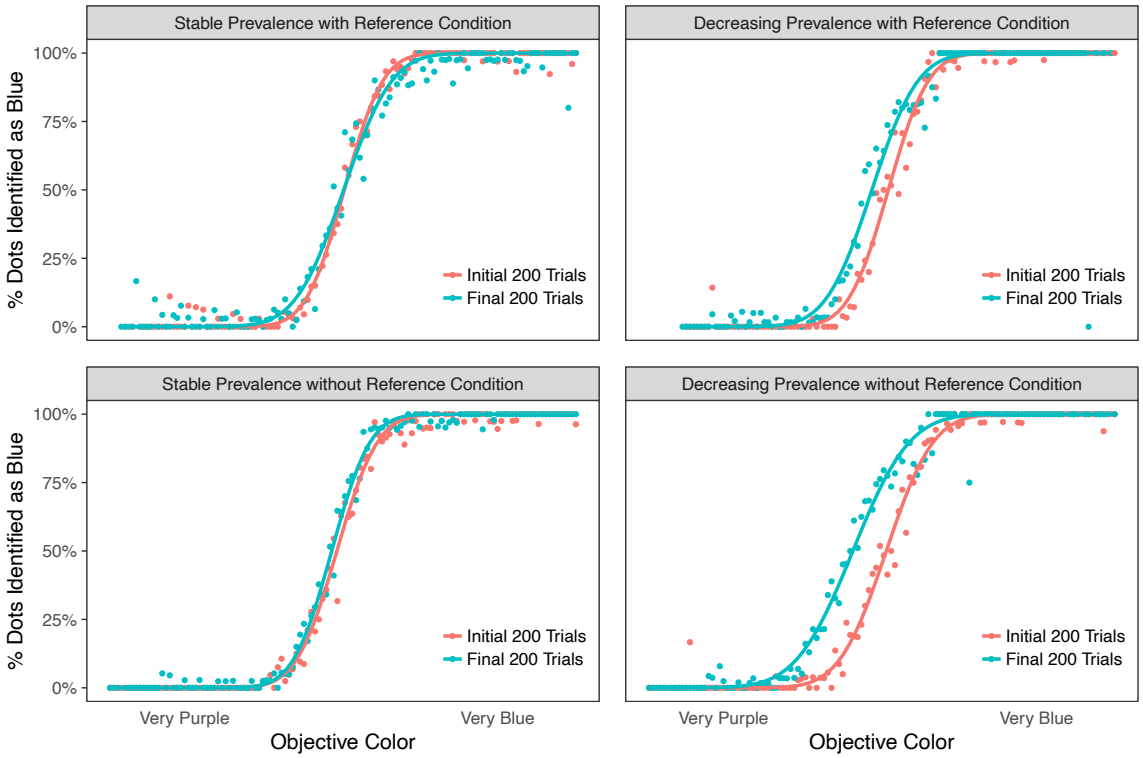
condition, b = 11.91, SE = 1.93, z = 6.16, p < 0.001, *95% CI* [8.12, 15.70], and the *decreasing + no reference* condition, b = 19.97, SE = 1.96, z = 10.18, p < 0.001, *95% CI* [16.13, 23.82].

The *decreasing + reference* condition differed significantly from the *decreasing + no reference* condition, b = 4.91, SE = 2.00, z = 2.5, p = 0.03, *95% CI* [0.99, 8.84], but not from the *stable + no reference* condition, b = -3.18, SE = 1.95, z = -1.6, p = 0.11, *95% CI* [-6.98, 0.66].

Finally, the *stable + no reference* condition differed significantly from the *decreasing + no reference* condition, b = 8.08, SE = 2.09, z = 3.9, p < 0.001, *95% CI* [3.99, 12.17]. (All reported p-values are adjusted for multiple comparisons using the Holm correction).

Figure 7 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-200) and the final trials (600-800). The positive slope of all curves indicates that in all conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But the panels differ in an important way. The two curves in each of the two leftmost panels are largely superimposed, indicating that participants in the stable condition were just as likely to identify a dot as blue when it appeared on a final trial as when it appeared on an initial trial, whether or not a visual reference was present. But in the two rightmost panels, the two curves are offset in the middle, indicating that participants in these decreasing conditions were more likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial, whether or not a visual reference was present. In short, when blue dots became less prevalent, participants identified as blue some dots that they had earlier identified as not blue, and they did this even when a visual reference was present to remind them of their original threshold value between blue and purple.

FIGURE 7: Results for Study 6



The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

Conclusion

Participants who experienced a decrease in the prevalence of blue dots 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, even when the trial under evaluation was accompanied by a maximally ambiguous stimulus representing the participant's personal threshold at the start of the study. As to whether the presence of a visual reference change the magnitude of the prevalence shift (even if it did not eliminate it entirely), there was a small, significant difference in the odds of identifying a dot as blue in the decreasing condition, depending on whether a visual reference

was present. Specifically, participants were slightly less likely to identify dots as blue in later trials (after the prevalence had decreased) when a visual reference was present. This is tentative evidence that the presence of a visual reference, while unable to prevent observers from shifting their thresholds for the color blue in response to a prevalence decrease, may slightly reduce the size of the effect.

Studies 1-6 show that people respond to a change in the prevalence of a color by shifting their conceptual boundaries of that color, even with explicit motivations or instructions not to do so, and regardless of the speed of the shift or the presence of a visual reference. Does this phenomenon generalize from the simple perception of color to more complex judgments? Studies 7 and 8 were designed to see if prevalence also impacted judgments of facial threat and ethics in scientific research, respectively.

STUDY 7: FACIAL THREAT

Overview

I showed participants in Study 7 a series of human faces on a computer screen and asked them to determine whether the person they saw (hereinafter referred to as *the target person*) was a threat or was not a threat. Over the course of many trials, I decreased the prevalence of threatening target persons for some participants. I predicted that these participants would respond to the decreasing prevalence of threatening target persons by identifying some target persons as threats whom they had previously identified as non-threats.

Methods

Sample. Participants were 49 students at Harvard University (28 male, 20 female, and 1 gender unspecified, M_{age} = 20.8 years, SD = 2.0 years) who received either money or course

credit in exchange for their participation. One male participant reported having a form of prosopagnosia (face blindness), and his data were excluded, leaving 48 participants in the data set.

Procedure. Upon arrival at the laboratory, participants were escorted to a room equipped with a computer display and keyboard, and they remained there for the duration of the study. Participants were told that a series of target persons would appear on the screen, one at a time, and that their task was to decide whether each target person was or was not a threat, and to indicate their decision by pressing one of two keys on the keyboard that were respectively labeled "threat" and "no threat." On each trial, a computer-generated image of a target person's face appeared on a solid gray background. To generate these images, I took the most and least threatening faces from a pre-scaled series of computer-generated faces created by Todorov and colleagues (2013; 2011) and then used Fantamorph (http://www.fantamorph.com/) to incrementally morph the faces into one another. This produced a continuum of 60 target persons whose facial expressions ranged from not very threatening to very threatening. Sample faces are shown in Figure 8.

FIGURE 8: Examples of Faces Used in Study 7

The target person continuum ranged from 1 (not threatening) to 60 (very threatening) and this figure shows (from left to right) faces 1, 10, 20, 30, 40 50, and 60. The four target persons on the left are from the no threat continuum and the three target persons on the right are from the threat continuum.

Although the threateningness of a face is inherently subjective, for the sake of consistency I refer to the mean rating of each target as its *objective threateningness*. Each target 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 threatening targets might vary over blocks. Participants completed 10 practice trials to ensure that they understood the procedure, and then completed 800 test trials. To help participants remain attentive, I allowed them to take a break every 50 trials.

I created two conditions by dividing the target continuum into two halves that I will refer to as the "no threat continuum" and the "threat continuum." Half the participants were randomly assigned to the *stable* condition. In this condition, I determined the threateningness of the target shown on each trial by randomly sampling the two continua with equal probability. I will refer to the probability that a target was sampled from the threat continuum 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 continua 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 (see Appendix II).

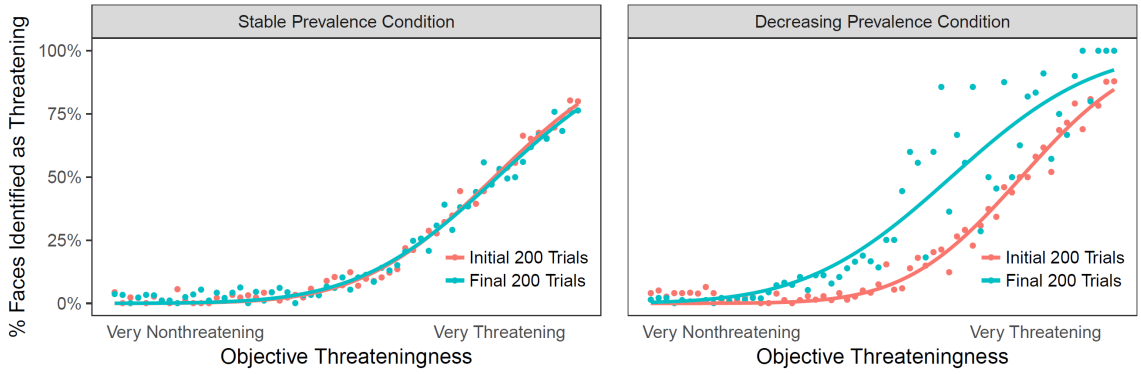
Results

Did the decrease in the prevalence of threatening targets cause participants' concepts of *threat* to expand*?* To find out, I fit a binomial generalized linear mixed model to my data in R using the lme4 package. The dependent variable was the participant's *identification* of a target as threatening or not threatening. The independent between-participants variable was the participant's *condition* (stable or decreasing). The independent within-participants variables were (a) the target's position on the continuum or what I will call its *objective threateningness* (which ranged from 0% threatening to 100% threatening) and (b) the *trial number* (which ranged from 1 to 800). I included condition, trial number, and objective threateningness (and all interactions between them) as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. The inclusion of random intercepts significantly improved model fit relative to the baseline model, $\chi^2(2) = 649.03$, $p < 0.001$, as did the inclusion of random slopes, $\chi^2(2) = 974.24$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective threateningness significantly improved model fit, $\chi^2(1) = 32.24$, $p < 0.001$. The generalized linear mixed model revealed that a Condition X Objective Threateningness X

Trial Number interaction predicted participants' identifications, $b = 4.84$, $SE = 0.86$, $z = 5.61$, *95% CI* [3.12, 6.53], $R^2_{GLMM(c)} = 0.75$.

Figure 9 shows the percentage of targets at each point on the continuum whom participants identified as a threat on the initial 200 trials and on the final 200 trials. Participants in the stable condition were just as likely to identify a target as a threat when that target appeared on a final trial as when that target appeared on an initial trial, but participants in the decreasing condition were more likely to identify a target as a threat when the target appeared on a final trial than when the target appeared on an initial trial. In other words, when the prevalence of threatening targets decreased, participants' concepts of *threat* expanded to include targets that it had previously excluded.

FIGURE 9: Results for Study 7



The x-axes show the target's objective threateningness (as determined by raters) and the y-axes show the percentage of trials on which participants identified that target as a threat.

Conclusion

Participants who experienced a decrease in the prevalence of threatening faces were more likely to identify faces as threatening when those faces appeared on a final trial than when those faces appeared on an initial trial, similar to the shifts seen in color identification in the preceding studies.

Facial emotion perception, like color perception, is visual in nature. Study 8 was designed to see is a similar effect of prevalence on identification can be found in a nonvisual, socially relevant domain.

STUDY 8: ETHICAL DECISIONS

Overview

In Study 8, I showed participants a series of proposals for scientific studies and asked them to decide whether researchers should be prohibited from conducting the study or should be allowed to conduct the study. The proposals varied in their ethicality. Over the course of many trials, I decreased the prevalence of unethical proposals for some participants. I predicted that these participants would respond to the decrease in the prevalence of unethical proposals by rejecting some proposals that were ethically identical to those they had previously accepted.

Methods

Whereas colors and computer-generated faces vary on physical continua that can be measured on a ratio scale, ethicality can at best be measured on an ordinal scale. As such, the materials and procedures for Study 8 differed somewhat from the materials and procedures used in my previous studies.

Materials. Together with a team of research assistants, I wrote 381 short proposals for scientific experiments involving human participants. The proposals contained between 5 and 37 words (M = 25.34 words). We used our own judgment to preliminarily classify each proposal as either ethical, ambiguous, or unethical. I then recruited 361 U. S. residents (198 male, 161 female, 2 gender unspecified) via Amazon Mechanical Turk and asked them to read and rate a subset of these proposals. I will refer to these participants as *the raters*. Raters were told that (a) the proposals described experiments that were designed to be conducted with adults who had volunteered to take part in exchange for money; (b) all the studies described in the proposals were research on human behavior; (c) when scientists lie to participants either before or during a study, they always tell those participants the truth when the study is over; and (d) participants are always free to withdraw from a study at any time.

Each rater was paid $1 to read and rate 76 proposals. I divided the 381 proposals into a set of 15 proposals that were seen by all raters (the constant set) and a set of 366 proposals that were seen by a subset of raters (the variable set). Specifically, the 366 proposals were divided into 6 sets of 61 proposals (the variable sets), and each rater saw one of these 6 variable sets as well as the constant set of 15 proposals. Twenty-one of the proposals in each of the variable sets had been preliminarily classified as ethical, 23 had been preliminarily classified as ambiguous, and 17 had been preliminarily classified as unethical. The 61 proposals in each of the variable sets were presented in random order, and after the 20th and 40th, and 61st proposals I included a "catch question" to ensure that raters were reading carefully (viz., "If you're actually reading this question, please select the number 3 as your response. Thank you for reading all the questions carefully"). Each rater first saw one of the 6 variable sets of 61 proposals, and then saw the 15 proposals in the constant set. After seeing each proposal, raters were asked the question "Should

this experiment be allowed to be conducted?" which they answered using a 7-point Likert scale whose endpoints were anchored with the phrases "Definitely not" (1) and "Definitely" (7). Raters spent between 3.18 and 53.72 minutes (M = 16.09 min) making their ratings. After they did so, raters completed several other measures including a Turing test (e.g., "If you're reading this, type the word *banana*"), and supplied demographic information.

I excluded the ratings of two male and three female raters who failed the Turing test, and then computed the mean rating of each proposal. Despite the fact that participants' ratings were inherently subjective, for the sake of consistency I will refer to the mean of each proposal's ratings as its *objective ethicality.* Each rater saw 76 proposals. Fifteen of these proposals (the constant set) were seen by all raters, which allowed us to estimate how much the complete pool of raters agreed with regard to judgments of ethicality. Inter-rater reliability was quite high (Cronbach's alpha = .85), indicating that raters were in very close agreement about the objective ethicality of the proposals. I used each proposal's objective ethicality to classify it as a member of one of three categories. To ensure that I had a sufficient number of proposals in each of these categories, I classified proposals whose objective ethicality was greater than 6 and less than or equal to 7 as *ethical*; proposals whose objective ethicality was greater than 4 and less than or equal to 6 as *ambiguous*; and proposals whose objective ethicality was less than or equal to 4 and greater than or equal to 1 as *unethical*. I then selected the proposals in each of the three categories whose objective ethicality ratings had the lowest standard deviations. Specifically, I selected 113 *ethical proposals* (e.g., "Participants will make a list of the cities they would most like to visit around the world, and write about what they would do in each one"), 80 *ambiguous proposals* ("Participants will be given a plant and told that it is a natural remedy for itching. In reality, it will cause itching. Their reaction will be recorded"), and 80 *unethical proposals* (e.g.,

"Participants will be asked to lick a frozen piece of human fecal matter. Afterwards, they will be given mouthwash. The amount of mouthwash used will be measured"). These 273 proposals were used as materials in Study 8.

Sample. Participants in Study 8 were 84 students at Harvard University (16 male, 66 female, 2 gender unspecified, M_{age} = 20.73 years, SD = 2.8 years) who received either money or course credit for their participation.

Procedure. Upon arrival at the laboratory, participants were escorted to a room equipped with a computer display and keyboard, and they remained there for the duration of the study. Participants were told that a series of proposals for scientific studies would appear on the screen, one at a time, and that their task was to decide whether researchers should or should not be allowed to conduct each study. They were asked to indicate their decision about each proposal by pressing one of two keys on the keyboard that were respectively labeled "approve" and "reject." On each trial, participants read one of 273 proposals. Each proposal appeared on the screen and remained there until participants pressed one of the response keys. Participants were told that there would be 240 trials divided into 10 blocks, and that the ethicality of the proposals might vary over blocks. Participants completed one practice trial to ensure that they understood the procedure, and then completed 240 test trials. To help participants remain attentive, I allowed them to take a break every 24 trials.

I created two conditions. Half the participants were randomly assigned to the *stable* condition. In this condition, I determined the ethicality of the proposal on each trial by randomly sampling the three ethicality categories (ethical, ambiguous, and unethical) with equal probability. I will refer to the probability that a proposal was sampled from the unethical category as the *signal prevalence*. In the stable condition, the signal prevalence on trials 1-240

was 33.3%. In the *decreasing* condition, I sampled the three categories with unequal probability on some trials. Specifically, in the decreasing condition, the signal prevalence was 33.3% on trials 1-96; 25% on trials 97-120; 16.6% on trials 121-144; 8.3% on trials 145-168; and 4.12% on trials 169-240. After completing the identification task, participants completed a questionnaire asking some basic demographics and their impressions of the task (see Appendix II).

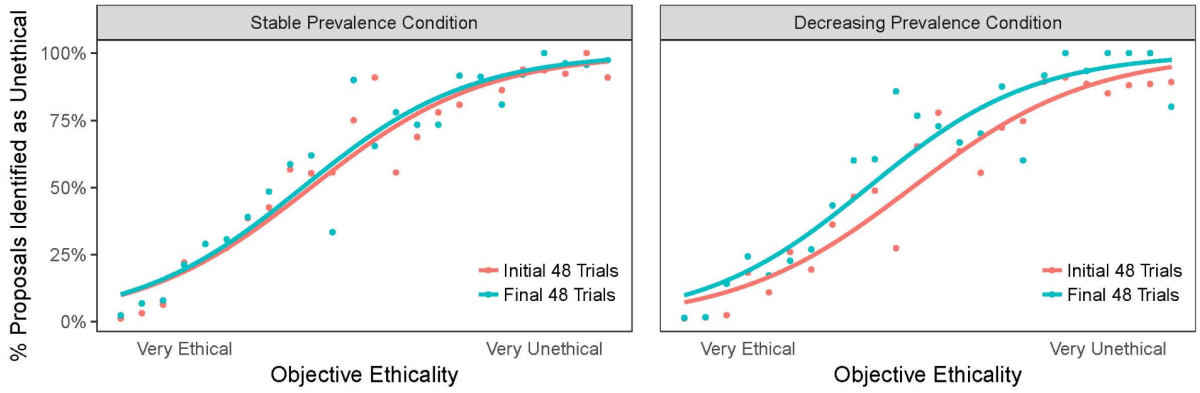
Results

Did the decrease in the prevalence of unethical proposals cause participants' concepts of *unethical* to expand? To find out, I fit a binomial generalized linear mixed model to my data in R using the lme4 package. The dependent variable was a binary measure of whether a proposal was accepted or rejected. The independent between-participants variable was the prevalence of unethical proposals (stable or decreasing), and my independent within-participants variables were (a) the trial number (which ranged from 1 to 240) and (b) the objective ethicality rating of each proposal, which were reverse-scored for analysis so that 1 = "This experiment should definitely not be allowed" and 1 = "This experiment should definitely be allowed". I included prevalence, trial number, and objective ethicality rating as fixed effects in my model, along with all interactions. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. Model fit was significantly improved by both random slopes for trial, $\chi^2(2) = 63.69$, $p < 0.001$, and random intercepts for participants, $\chi^2(2) = 404.51$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective ethicality rating significantly improved model fit, $\chi^2(1) = 24.71$, $p < 0.001$.

The generalized linear mixed model revealed that a Prevalence X Objective Ethicality Rating X Trial Number interaction predicted participants' identifications, $b = -5.19$, $SE = 1.04$, z

= -4.97, *95% CI* [-7.15, -3.16], $R^2_{GLMM(c)}$ = 0.73. Figure 10 shows the percentage of proposals that participants rejected on the initial 48 trials and on the final 48 trials. Participants in the stable condition were just as likely to reject ethically ambiguous proposals that appeared on a final trial and on an initial trial, but participants in the decreasing condition were more likely to reject ethically ambiguous proposals that appeared on a final trial than on an initial trial. In other words, when the prevalence of unethical research proposals decreased, participants' concept of *unethical* expanded to include proposals that it had previously excluded.

FIGURE 10: Results for Study 8



The x-axes show the proposal's objective ethicality (as determined by raters) and the y-axes show the percentage of trials on which participants rejected the proposal.

Conclusion

Participants who experienced a decrease in the prevalence of unethical study proposals were more likely to identify study proposals as unethical when those proposals appeared on a final trial than on an initial trial, similar to the shifts seen in color and threat identification in the

preceding studies. This suggests that the phenomenon seen in Studies 1-7 is not limited to judgments of visual stimuli, though whether or not the same computational mechanism produces the behavior seen across these three domains is unclear.

STUDY 9: COLOR AND ANIMACY JUDGMENTS

Overview

Studies 1-8 all gave participants the subjective experience of the overall rate of detection either dwindling or increasing dramatically over time. In other words, participants in conditions where the prevalence shifted started and ended the experiment making identifications of a particular concept very different rates. Is this a necessary condition for prevalence-induced concept change? In this account, the uncertainty or heightened vigilance prompted by stimuli becoming rare or common would lead to decision threshold shifts.

A useful method to test this account was demonstrated by Wolfe and colleagues (Wolfe et al., 2007) (see also Hout et al., 2015) who devised a detection task to study the Low-Prevalence Effect in which one class of signals was very prevalent while others were rare. This left overall target prevalence at a constant 50% while still allowing for analysis of trials with very rare targets. Applying this design to my phenomenon, I would predict that observers who are frequently finding instances from one kind of concept (even if others dwindle) might not shift their thresholds.

In Study 9, I replicated the procedure for Study 1, except that in Study 9, trials where participants evaluated dots to decide whether or not they were blue were interleaved with trials in which participants evaluated faces to decide if they were animate. The prevalence change in blue dots was always accompanied by the opposite change in prevalence of animate faces, with the

goal of keeping the overall rate of instance detection and button-pressing relatively stable over time.

Methods

Sample. Participants were 108 students at Harvard University (28 males, 80 females, M_{age} = 19.24 years, SD = 1.40 years) who received course credit in exchange for their participation.

Procedure. The method for Study 9 was similar to the methods for Study 1, except that there were two types of trials in the task — color trials and animacy trials. Color trials were identical to trials from Study 1. In animacy trials, participants saw an image of a face taken from stimuli created by Looser & Wheatley (2010). On each trial, an image of a target person's face appeared on a solid gray background. The faces incrementally morphed between two photographs – one photograph of a human face, and one of a doll face that looked similar to that human face. This process produced a continuum of 101 target persons whose facial animacy ranged from not at all animate to very animate. Sample faces are shown in Figure 11.

FIGURE 11: Examples of Faces Used in Study 9

The target person continuum ranged from 1 (inanimate) to 101 (animate) and this figure shows (from left to right) faces 1, 20, 40, 60, 80, and 101. The four target persons on the left are from the inanimate continuum and the three on the right are from the animate continuum.

There were three conditions in the study. In the *stable* condition, the prevalence of both blue dots and animate faces remained at 50% throughout the task (trials 1-960). In the *decreasing blue*/*increasing animacy* condition, the prevalence of blue dots was 50% on trials 1-240, 40% on trials 241-300, 28% on trials 301-360, 14% on trials 361-420, and 6% on trials 421-960. The prevalence of animate faces was 50% on trials 1-240, 60% on trials 241-300, 72% on trials 301-360, 86% on trials 361-420, and 94% on trials 421-960. Conversely, in the *increasing blue/decreasing animacy* condition, the prevalence of blue dots was 50% on trials 1-240, 60% on trials 241-300, 72% on trials 301-360, 86% on trials 361-420, and 94% on trials 421-960. The prevalence of animate faces was 50% on trials 1-240, 40% on trials 241-300, 28% on trials 301-360, 14% on trials 361-420, and 6% on trials 421-960. After completing the identification task, participants completed a questionnaire asking some basic demographics and their impressions of the task (see Appendix II).

Results

Did the decrease in the prevalence of blue dots cause participants' concepts of blue to contract (rather than to expand), even when accompanied by the increase in prevalence of a second, unrelated stimulus (animate faces)? To find out, I fit separate models to my data to analyze participant responses to colors and faces.

To analyze color trials, I fit a binomial generalized linear mixed model to my data in R using the lme4 package. 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, decreasing blue/increasing animacy, or increasing blue/decreasing animacy). The independent within-participants variables were (a) the dot's RGB value or what I will call its objective color (which ranged from 0% blue to 100% blue) and (b) the trial number (which

ranged from 1 to 960). I included condition, trial number, and objective color (and all interactions between them) as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. The inclusion of random slopes significantly improved model fit relative to the baseline model, $\chi^2(2) = 290.63$, $p < 0.001$, as did the inclusion of random intercepts, $\chi^2(2) = 966.61$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective color significantly improved model fit, $\chi^2(2) = 48.30$, $p < 0.001$, $R^2_{GLMM(c)} = 0.87$.

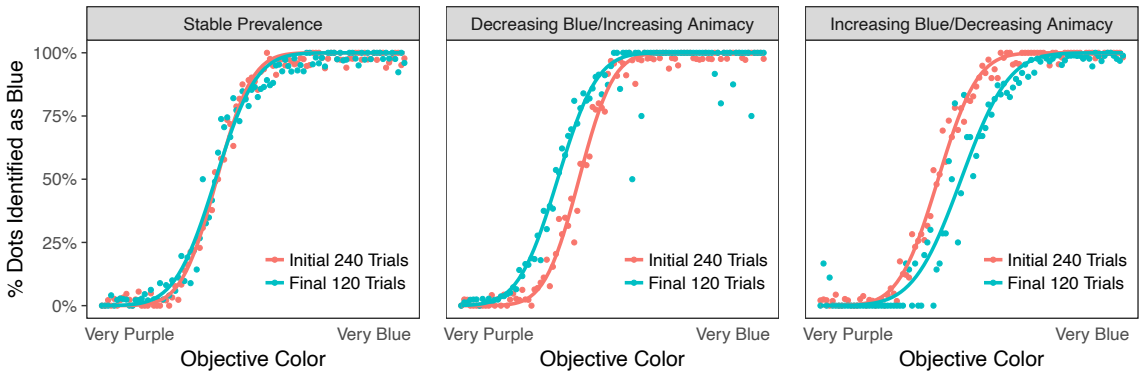
To analyze animacy trials, I fit a second binomial generalized linear mixed model to my data. The dependent variable was the participant's identification of a face as animate or inanimate. The independent between-participants variable was the participant's condition (stable, decreasing blue/increasing animacy, or increasing blue/decreasing animacy). The independent within-participants variables were (a) the face's value on the animacy morph spectrum or what I will call its objective animacy (which ranged from 0% animate to 100% animate) and (b) the trial number (which ranged from 1 to 960). I included condition, trial number, and objective animacy (and all interactions between them) as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. The inclusion of random slopes significantly improved model fit relative to the baseline model, $\chi^2(2) = 1149.2$, $p < 0.001$, as did the inclusion of random intercepts, $\chi^2(2) = 910.68$, $p < 0.001$. Additionally, the inclusion of the three-way interaction between condition, trial number, and objective animacy significantly improved model fit, $\chi^2(2) = 35.39$, $p < 0.001$, $R^2_{GLMM(c)} = 0.83$.

For color trials, the generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications of colors. Specifically, the *stable* condition differed from the *decreasing blue/increasing animacy* condition, $b = 6.20$, $SE = 1.25$, $z = 4.95$, $p < 0.001$, *95% CI* [3.74, 8.66], as well as from the *increasing blue/decreasing animacy* condition, $b = -3.58$, $SE = 1.19$, $z = -3.01$, $p = 0.002$, *95% CI* [-5.91, -1.25]. Finally, the *decreasing blue/increasing animacy* condition also differed significantly from the *increasing blue/decreasing animacy* condition, $b = -9.78$, $SE = 1.30$, $z = -7.50$, $p < 0.001$, *95% CI* [-12.33, -7.22].

Figure 12 shows the percentage of dots of each color that participants in each condition identified as blue on the initial trials (1-1240) and the final trials (721-960). The positive slope of all curves indicates that in both conditions, participants' identifications were highly correlated with the dot's position on the color spectrum. But the three panels differ in an important way. 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 a final trial as when it appeared on an initial trial. But the two curves in the middle and right panels are offset in the middle, indicating that in the presence of a decrease in the prevalence of blue dots, participants were more likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial, while an increase in the prevalence of blue dots produced the opposite pattern, making participants less likely to identify dots from the middle of the color spectrum as blue when those dots appeared on a final trial than when they appeared on an initial trial. In short, when blue dots became less prevalent, participants identified as blue some dots that they had earlier identified as not blue, even when animate faces were increasing in prevalence at the same time. Conversely, when blue dots

became more prevalent, participants identified as not blue some dots that they had earlier identified as blue, even when animate faces were decreasing in prevalence at the same time.

FIGURE 12: Results for Study 9 Color Trials



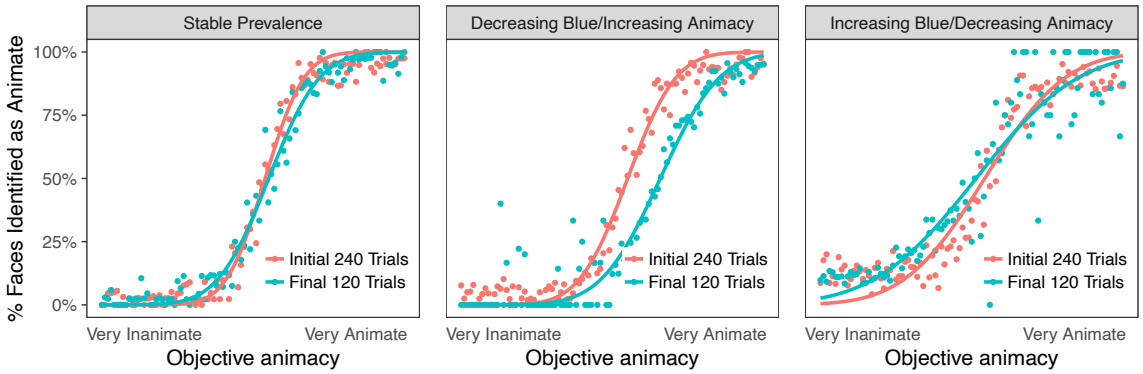
The x-axes show the dot's objective color (i.e., its location on the spectrum) and the y-axes show the percentage of trials on which participants identified that dot as blue.

On animacy trials, the generalized linear mixed model revealed that a Condition X Objective Color X Trial Number interaction predicted participants' identifications of animacy. Specifically, the *stable* condition differed from the *decreasing blue/increasing animacy* condition, $b = 6.26$, $SE = 1.14$, $z = 5.50$, $p < 0.001$, *95% CI* [4.03, 8.49], but not from *increasing blue/decreasing animacy* condition, $b = 1.42$, $SE = 0.94$, $z = 1.51$, $p = 0.13$, *95% CI* [-0.42, 3.26]. Finally, the *decreasing blue/increasing animacy* condition differed significantly from the *increasing blue/decreasing animacy* condition, $b = -4.84$, $SE = 0.91$, $z = -5.29$, $p < 0.001$, *95% CI* [-6.63, -3.05].

Figure 13 shows the percentage of faces at each point on the continuum that participants in each condition identified as animate on the initial trials (1-240) and the final trials (721-960).

The positive slope of all curves indicates that in both conditions, participants' identifications were highly correlated with the face's position on the animacy continuum. But the three panels differ in an important way. The two curves in the left and right panels are nearly perfectly superimposed, indicating that participants in the stable and increasing blue/decreasing animacy conditions were just as likely to identify a face as animate when it appeared on a final trial as when it appeared on an initial trial. But the two curves in the middle panel are offset in the middle, indicating that in the presence of an increase in the prevalence of animate faces, participants were more likely to identify faces from the middle of the animacy continuum as animate when those faces appeared on a final trial than when they appeared on an initial trial. In short, when animate faces became more prevalent, participants identified as inanimate some faces that they had earlier identified as animate, even when blue dots were decreasing in prevalence at the same time. However, when animate faces instead decreased in prevalence, participants did not call a wider range of faces, even when blue dots were increasing in prevalence at the same time.

FIGURE 13: Results for Study 9 Animacy Trials



The x-axes show the face's objective animacy (i.e., its location on the continuum) and the y-axes show the percentage of trials on which participants identified that face as animate.

As a manipulation check, I fit a third binomial generalized linear mixed model to my data, collapsing across animacy and color trials to see whether overall response rates remained stable across conditions. The dependent variable was the participant's identification of either a face as animate or inanimate or a color as blue or not blue. The independent between-participants variable was the participant's condition (stable, decreasing blue/increasing animacy, or increasing blue/decreasing animacy). The independent within-participants variable was the trial number (which ranged from 1 to 960). I included condition and trial number as fixed effects in my model. I included as random effects (a) intercepts for participants (who may have entered my study with different thresholds) and (b) slopes for trial number. The inclusion of random slopes significantly improved model fit relative to the baseline model, $\chi^2(2) = 304.54$, $p < 0.001$, as did the inclusion of random intercepts, $\chi^2(2) = 295.62$, $p < 0.001$. The inclusion of the two-way interaction between condition and trial number did not significantly improve model fit, $\chi^2(2) = 4.40$, $p < 0.11$, $R^2_{GLMM(c)} = 0.05$.

The generalized linear mixed model revealed no evidence of a Condition X Trial Number interaction predicting participants' identifications of animacy and color. Specifically, the *stable* condition did not differ from the *decreasing blue/increasing animacy* condition, b = -0.11, SE = 0.13, z = -0.87, p = 0.76, *95% CI* [-0.37, 0.14], nor from the *increasing blue/decreasing animacy* condition, b = 0.15, SE = 0.13, z = 1.20, p = 0.70, *95% CI* [-0.10, 0.40]. Additionally, the *decreasing blue/increasing animacy* condition did not differ significantly from the *increasing blue/decreasing animacy* condition, b = 0.26, SE = 0.13, z = 2.10, p = 0.14, *95% CI* [0.02, 0.51]. Figure 14 shows the proportion of stimuli identified as either blue or animate as a function of trial number in the task, split by the three conditions of the study. The three lines are nearly perfectly superimposed, indicating that participants identified roughly the same proportion of stimuli as blue or animate through the task, whether or not blue dots or animate faces changed in prevalence in any way.

FIGURE 14: Results for Study 9 Response Rates

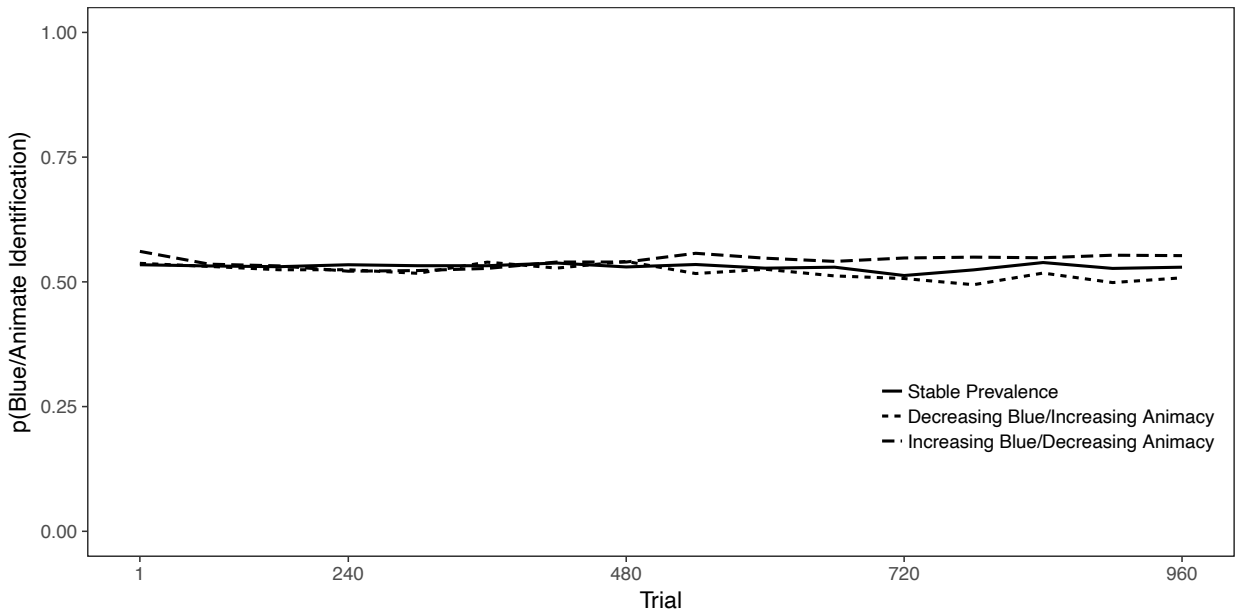


FIGURE 14 (Continued).

The x-axis shows the trial in the task (from 1-960) and the y-axis shows the percentage of trials on which participants pressed the right-hand response button, either identifying a dot as blue or a face as animate. The three lines show the three conditions of the study.

Conclusion

Participants who experienced a decrease in the prevalence of blue dots 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, even when these trials were interleaved with facial animacy stimuli that were increasing in prevalence. In other words, the fact that roughly half of the total stimuli were being identified with each input button throughout the task, regardless of the current prevalence, did not eliminate the effect in color perception. The manipulation check collapsing across color and animacy trials suggests that the study design was successful in helping participants maintain stable rates of overall stimulus identification in the face of any prevalence shifts.

Animacy trials did not exhibit the same conceptual shift as color trials when the prevalence of animate faces was decreased, though the predicted shift was observed when the prevalence was instead increased. Why were decreasing prevalence trials unsuccessful in eliciting a threshold shift? One possibility is that participants were unable to expand their threshold of animate faces to include more faces as animate beyond a certain limit. In other words, while animacy judgments may be in part subjective (like color judgments), there may remain hard perceptual limits on what range of faces could ever be called animate by most observers.

INTERIM DISCUSSION

The nine experiments presented here explored the tendency for human observers to respond to changes in the prevalence of a concept by shifting their thresholds for detecting instances of that concept. Specifically, as conceptual instances were made rarer, observers in a range of domains relaxed their thresholds for what counted as an instance of that concept, while as instances became more common, thresholds were instead tightened. This phenomenon may have serious social consequences, since in many detection tasks, thresholds should not be influenced by prevalence. Several attempts to debias participants (Studies 3-6, 9) were mostly unsuccessful. Understanding the underlying computational process involved in this phenomenon may enable the development of more effective interventions to prevent it. Why do humans shift their detection thresholds in response to prevalence changes, and why in the direction documented in these studies? I will review several potential explanations.

Demand effects

Is it possible that my participants, when faced with a prevalence change, consciously and purposefully shifted their criteria in response to that change? I will describe a few different versions of this account, and argue that all of them are unlikely to explain my findings.

Could participants have inferred that consciously shifting their decision thresholds was an appropriate or expected response to prevalence changes? I made efforts to remove any such incentives in the task, overt or implied. Participants were told that their judgments had no objectively right or wrong answers. They were also explicitly instructed to disregard any prevalence changes they did notice. Additionally, with the exception of Study 4, participant compensation remained the same regardless of how they responded in the task.

Could participants have been unaware of the prevalence change as it happened, or confused by it once it did? Perhaps participants assumed that approximately 50% of stimuli

would belong to each concept. In this account, when the prevalence did change, participants were confused, and tried to consciously maintain their previous rates of detection by relaxing their decision thresholds, because they misunderstood the rules of the task or were misled by the experimenter. Study 3 presents evidence counter to this confusion account, because participants were explicitly forewarned about a prevalence change, and still relaxed their thresholds in response, even when they were warned that it would occur. Alternatively, the gradual nature of the prevalence change may have made it hard to notice. Study 5 speaks to this concern – participants still exhibited the shift even when the prevalence decrease was abrupt, and thus, more noticeable.

Could participants in my studies have grown more adept at distinguishing between concepts over time? In this account, my phenomenon is not a shifting in detection standards based on some sort of perceptual bias, but rather, reflects participants' improved skill at locating the concept they are searching for. For example, participants in Study 1 may have actually become more sensitive to the presence of blue hue in dots as the study went on. This is known as perceptual learning (Goldstone, 1998), a process by which observers become more skilled at detecting stimuli with practice. However, if participants were becoming more skilled over time at detecting blue dots, threatening faces, or unethical experiments, then I should also have seen increased rates of detection in the control conditions of these tasks, when the prevalence stays stable. I didn't observe this result, so while it's possible that my participants are becoming slightly more skilled at distinguishing subtly different stimuli from one another, perceptual learning doesn't account for the overall pattern of threshold shifting documented here.

I find it likely that my participants are shifting their standards not as a conscious strategy, but as an automatic response to a changing environment. If this is the case, how would such a

process operate? In the following section, I will review relevant theories that could help provide an answer.

Bayesian models of perception

A simple model of Bayesian probability estimation would predict tightening of an observer's decision threshold in response to a prevalence decrease. As a given stimulus becomes rare, the observer's prior over the likelihood of seeing that stimulus would go down, biasing them against identifying it again. Could an observer using some form of Bayesian inference instead display behaviors in line with prevalence-induced concept change? Imagine a model in which the observer is trying to use some form of Bayesian updating to infer the average intensity of stimuli, and then using that average as a detection threshold. As the prevalence of that stimulus went down over time, this would result in the threshold growing more liberal towards that concept.[4]

Adaptive value coding and normalization

If prevalence decreases cause observers to expand concepts like morality or threat, could this be due to the same process underlying visual aftereffects? I would argue that the phenomena are more similar than they might appear. One popular study of aftereffects involves staring at distorted faces for several minutes (an adaptation block), followed by a block of undistorted faces. In a study examining how the threshold between the colors blue and purple would change if blue dots became less prevalent over time, one could start by exposing participants to a block of dots with a mean equidistant between the two colors (50% blue), followed by an extended

[4] One notable recent Bayesian model of perception (Wei & Stocker, 2015) incorporates the principle of efficient coding to argue that using priors to infer the intensity of current stimuli can, under the right circumstances, produce an "anti-Bayesian" repulsive effect. I do not discuss Wei and Stocker's model here, but some of the models I do focus on involve some form of efficient coding (see General Discussion).

period with a mean that is very purple (6% blue prevalence). A vision researcher might describe such a task as an extended, noisy adaptation period. In this account, the first 200 or so trials would serve to "adapt" observers to dots with an average hue that is very blue. The subsequent trials would then produce a repulsive aftereffect, in which blue dots are perceived as even more blue than they were before, due to the observer adapting to higher frequency of very purple dots. It is entirely possible that my phenomenon recruits processes related to visual adaptation and aftereffects, a possibility first raised by Vickers & Leary (1983). Indeed, work on adaptation of facial emotion (Hsu & Young, 2004) and attractiveness (Rhodes et al., 2003) suggests that my phenomenon, at least in the domain of facial threat classification (Study 7), could be drawing on the same basic cognitive processes.

What mechanism underlies adaptation and aftereffects in the brain? A promising candidate here is normalization (Heeger, 1992). Rather than describing a specific brain area or circuit, this term describes a neural computation which the brain uses to encode value in variety of systems. When a given neuron receives an input, that activity is divided by the pooled activity of similarly tuned neurons. Normalization has been used to explain phenomena in vision (Carandini, Heeger, & Movshon, 1997), adaptive gain control (Louie, Grattan, & Glimcher, 2011), and higher level decisions and valuations as well (Louie, Khaw, & Glimcher, 2013). This process has been argued to be a feature of the efficient coding of the brain's architecture (Carandini & Heeger, 2012), in which it is more economical for the brain to store the differences between values than the absolute values themselves.

Range-Frequency Theory

My findings are in line with the predictions of Range-Frequency Theory (Parducci, 1963, 1965), which was developed to describe the influence that particular distributions of stimuli can

have on individual judgments of those stimuli. As applied to categorical judgments (Parducci & Wedell, 1986), the theory argues 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 stimuli values present in the observer’s current context. In my paradigm, using Study 1 as an example, Range-Frequency Theory would predict that the subjective perception of a dot’s color is based on both the most blue and least blue color seen in the study (the range), and the prevalence of dots at each color value in the study (the frequency).

Like work on vigilance or the low-prevalence effect, RFT is often used to model the behavior of observers either at a low frequency or a high frequency, but rarely on those undergoing the transition between these states. The prevalence shifts in my studies would manipulate what RFT would call the frequency parameter (how often instances of a concept appear). By contrast, the range of possible stimuli in my studies is fixed across experimental sessions.

A COMPUTATIONAL MODEL OF PREVALENCE-INDUCED CONCEPT CHANGE

Here I present a computational model of prevalence-induced concept change in color identification. The goal of the modeling is to posit several possible cognitive mechanisms for the phenomenon, and see which model best predicts the empirical data from Study 1. Aside from the basic model which serves as a control, each model I test invokes some way of using local prevalence to adapt the observer’s threshold between colors, or to change their subjective evaluation of color intensity itself.

Models

Model 1: Control CDF. As a control, I first implement an agent that explicitly has no ability to adapt subjective intensity or threshold values based on context. This basic model implements classic psychophysical categorical perception (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, 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:

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

The free parameter in the model is τ, the decision threshold.

Model 2: Bayesian adaptation model. The second model I test builds on the basic model by implementing dynamic updating of the decision threshold through Bayesian *maximum a posteriori* estimation (see 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_0, or the initial starting threshold and variance values that the observer

enters the task with. P(x_i|τ) is the likelihood of observing the current color given τ, in this case via a normal distribution with mean τ and deviation σ at value x_i. The entire posterior probability can be reformulated as follows:

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

(2.2)

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

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

(2.3)

This value is then passed to the normal CDF function as implemented in the basic observer model (equations 1.1 and 1.2), in order to classify the current color depending on whether it is greater or less than τ. The free parameter in this model is τ_0 (the initial decision threshold).

Model 3: Moving Window. This model uses the same normal CDF from Model 1, and attempts to adapt to local context by estimating 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:

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

(3.1)

where τ_i is the current threshold, τ_{i-1} is the estimated threshold from the previous trial, α is a scaling parameter (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 the basic model (equations 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) and τ_0 (the initial decision threshold).

Model 4: 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.

$$y_{ik} = w\left[\frac{x_i - x_{min,k}}{x_{max,k} - x_{min,k}}\right] + (1 - w)\left[\frac{rank_{ik} - 1}{n_k - 1}\right] \quad (4.1)$$

The subjective color value y_{ik} is then passed to the normal CDF function as implemented in the basic model (equations 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 number of trials included in the local context k), τ_0 (the decision threshold) and σ_0 (the standard deviation of the normal CDF). In my implementation, the range-weighting parameter w is fixed at .5.

Model 5: Range-only model. This model implements the range component of Range-Frequency theory from equation 4.1, with no tradeoff parameter (since there is no frequency component present).

$$y_{ik} = w\left[\frac{x_i - x_{min,k}}{x_{max,k} - x_{min,k}}\right] \quad (5.1)$$

The subject color value y_{ik} is then passed to the normal CDF function as implemented in the basic observer model (equations 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 number of trials included in the local context k) and τ_0 (the decision threshold).

Model 6: Frequency-only model. This model implements the frequency component of Range-Frequency theory from equation 4.1, with no tradeoff parameter (since there is no range component present).

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

(6.1)

The subject color value y_{ik} is then passed to the normal CDF function as implemented in the basic observer model (equations 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 number of trials included in the local context k) and τ_0 (the decision threshold).

Model 7: 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, times a scaling parameter α. The entire thing is then scaled by a scaling factor K, which represents gain.

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

(7.1)

The subject color value y_i is then passed to the normal CDF function as implemented in the basic observer model (equations 1.1 and 1.2), where it serves as the intensity of the current stimulus (x). The free parameters in the model are K (the gain scaling factor), n (the number of

trials included in the local context), α (the contextual scaling parameter) and τ_0 (the decision threshold).

Simulation and recovery

Data for four simulated agents in the color identification task from Study 1 (2 in the stable prevalence condition, and 2 in the decreasing prevalence condition) were generated in MATLAB from each of the seven models being tested. The true parameters for each agent were randomly sampled from uniform distributions using the bounds described at the end of this paragraph. Each agent completed 1000 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. Optimized parameters for each simulated agent were estimated with the MATLAB package *mfit* (Gershman, 2016). Five starting values were uniformly sampled for each parameter. In all seven models, τ_0 was sampled from the uniform distribution from 0.001 to 100, denoted as U(0.001,100). In the MAP model, σ_0 was sampled from U(0.01,100). In the Moving Window model, α was sampled from U(0.001,1). In the Range-Frequency model, σ_0 was sampled from U(0.01,100), and *n* was sampled from U(1,200). In the adaptive value coding model, *n* was sampled from U(1,200), K was sampled from U(1,100), and α was sampled from U(0.001,1). In the Range Only and Frequency Only models, *n* was sampled from U(1,200).[5]

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 seven models, fit in turn by each of the seven models (49 fittings in all). The protected exceedance probability is a

[5] Unless otherwise specified, in all models, σ (the standard deviation of the normal CDF) is fixed at 10. In the Range Frequency model, w (the range weighting parameter) is fixed at .5. Simulations with pilot data (not reported here) suggested that fixing these parameters led to better model performance. This may be because these parameters do not vary much between subjects, or because of the increased model complexity from adding an additional parameter, which is penalized in model selection.

value from Bayesian Model Selection (Rigoux, Stephan, Friston, & Daunizeau, 2014) that quantifies 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 BIC approximation of the marginal likelihoods of the fitted models. When responses simulated from each model were fit in turn by each of the seven models. Each model was best at fitting the responses generated by its own generative algorithm, with the exception of the Frequency Only model, which was outperformed by the control CDF model on its own data.

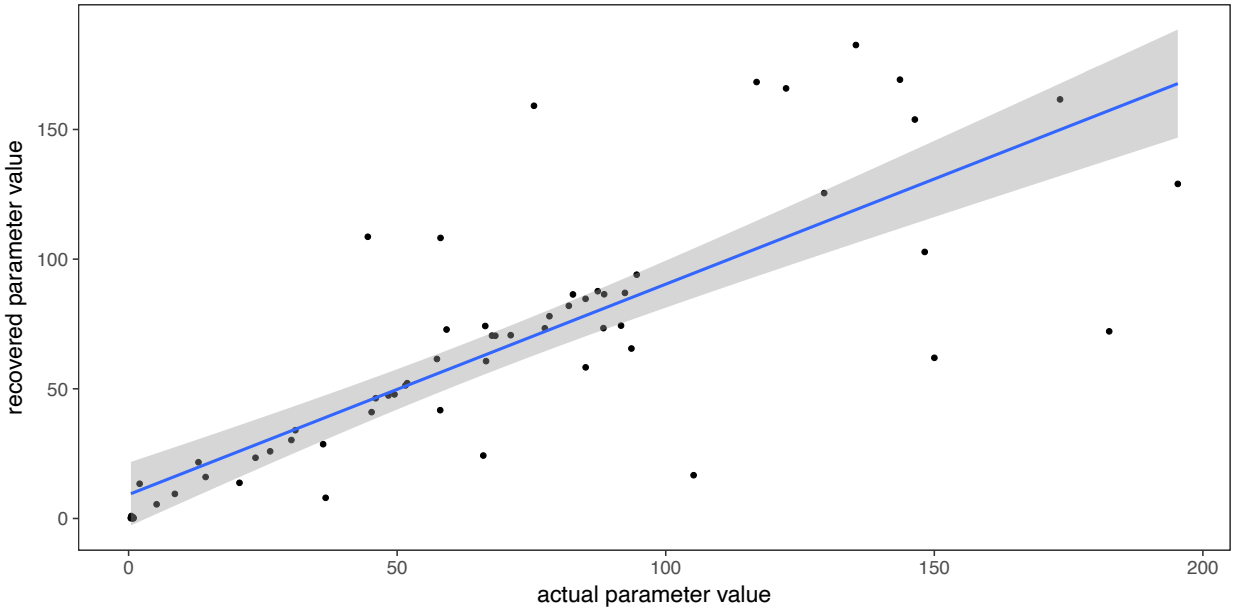
TABLE 1: Model selection for recovering simulated data

		Model that was used to fit responses						
		Control CDF	*MAP*	*Window*	*Range-Frequency*	*AVC*	*Range only*	*Frequency only*
Model that generated responses	*Control CDF*	0.787	0.035	0.036	0.035	0.035	0.035	0.036
	MAP	0.036	0.777	0.038	0.038	0.037	0.037	0.037
	Window	0.033	0.033	0.801	0.033	0.033	0.033	0.033
	Range-Frequency	0.034	0.034	0.034	0.798	0.034	0.034	0.033
	AVC	0.034	0.033	0.033	0.033	0.801	0.033	0.033
	Range only	0.065	0.065	0.064	0.146	0.065	0.530	0.064
	Frequency only	0.791	0.035	0.035	0.035	0.035	0.035	0.035

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

I also used each model to estimate the parameters for each of the four simulated participants. All models were reasonably accurate at estimating the true parameters of the simulated agents from that generative model ($r_{mean} = 0.82$). Figure 15 shows the relationship between the actual simulated parameters and the model-estimated parameters. Taken together with the Bayesian model selection on the simulated data, this result suggests that the predicted responses generated by at least six of the seven models tested here are distinct, and approximate recovery of not just those generative models, but the parameters of each simulated agent, is feasible.

FIGURE 15: Parameter recovery for simulated data



The x-axis shows the true parameter value of the simulated data, and the y-axis shows estimates of the same values recovered by the generative model that matches the true model.

Parameter estimation for actual human data

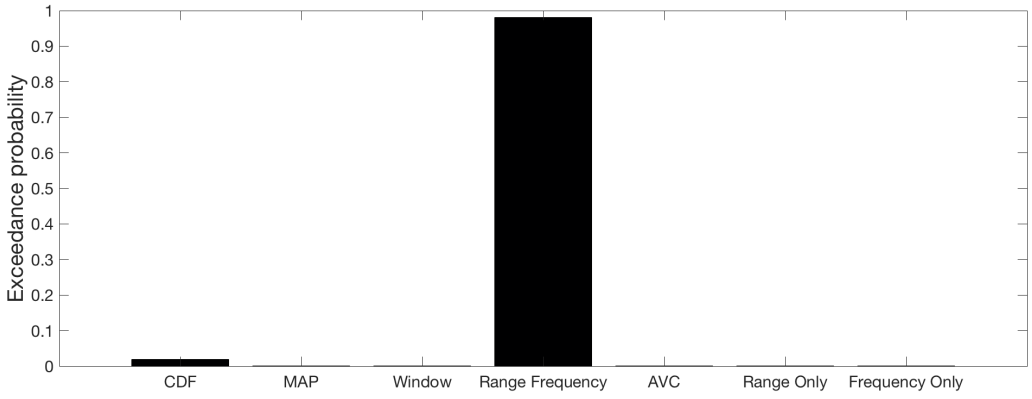
Optimized parameters for each human subject from Study 1 were again estimated using *mfit* (Gershman, 2016). Five uniformly sampled starting values were used for each parameter. The sampling distributions for each parameter were the same as in the simulation and recovery procedure described above. For all models, uniform priors were set on each parameter.

Model comparison for human subjects

I used Bayesian Model Selection (Rigoux et al., 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009) as implemented in the *mfit* package in order to compare the models to see which predictions best fit actual human data. Figure 16 shows the protected exceedance

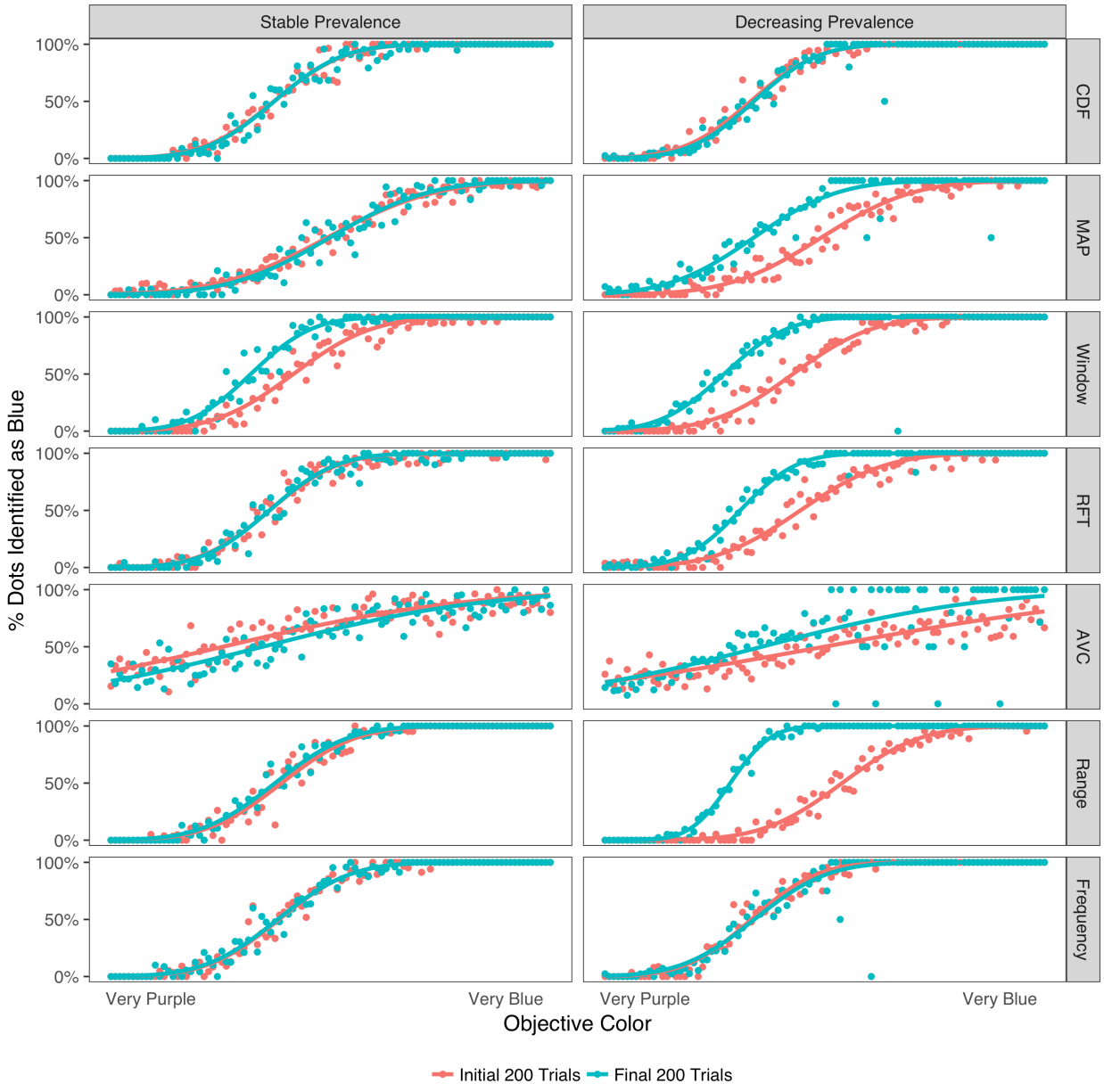
probabilities of the seven models. The Range-Frequency model strongly outperformed the other models in fitting the data, *pxp* = 0.98, *BOR* < 0.0001. Aggregating BIC values for each model produced the same pattern of results. Figure 17 shows the choice behavior of each model in the color identification task from Study 1. Unlike the control CDF model, the Range-Frequency model exhibits a shift in decision thresholds in the decreasing prevalence condition, similar to actual human subjects.

FIGURE 16: Bayesian Model Selection of Human Data from Study 1



Protected exceedance probability (PXP) scores (y-axis) are shown for color identification data from the 22 subjects in Study 1, fit by seven different models (x-axis).

FIGURE 17: Choice behavior of fitted models on a color identification task.



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 agents from each model identified that color as blue.

Conclusion

Several mechanisms from cognitive psychology and neuroscience could plausibly explain the phenomenon of prevalence-induced concept change. This work attempted to use trial-level computational modeling of human data to provide insight into which mechanism is most likely operating. My models suggest that computationally, prevalence-induced concept change in color identification operates in a process most similar to Range-Frequency Theory. Future research should investigate whether this model can also explain similar effects seen in other domains, such as facial threat perception and moral judgment. Ultimately, insight into the role of Range-Frequency Theory in high- and low-level perception and value judgments may suggest strategies to ameliorate the effect in situations where it is undesirable.

GENERAL DISCUSSION

The experiments recounted here document a tendency for human observers to respond to changes in concept prevalence by shifting their thresholds for detecting instances of that concept. Specifically, as instances were made rarer, observers in a range of domains relaxed their thresholds for what counted as an instance, while as they became more common, thresholds were instead tightened. Several attempts to debias participants (Studies 3-6 and 9) were unsuccessful. A computational model of the data from Study 1 suggested that the process that best characterized this phenomenon (at least, in color perception) was Range-Frequency Theory (Parducci, 1963).

The findings in these studies are in line not just with Range-Frequency Theory, but with several existing models and phenomenon described earlier, including visual aftereffects, divisive normalization, and sequential contrast effects. However, how can I account for the differences between these results and relevant findings from work on signal detection?

Reconciling with signal detection phenomena

Prevalence-induced concept change seems to be at odds with two well-known biases in signal detection, the vigilance decrement (Mackworth, 1948) and the Low-Prevalence Effect (Wolfe et al., 2007). I would argue that a few key differences in the tasks used to study these effects make them more reconcilable than they might seem at first.

Ground truth. In most signal detection research on vigilance or low prevalence, the tasks have some form of ground truth – an objective right or wrong answer for each decision the observer makes. For example, when evaluating whether an array of letters has any Ts, any given trial either has or doesn't have a T. This allows the researcher to classify every trial as either being signal-present or signal-absent and to know whether the observer responded correctly or incorrectly, which in turn allows for the calculation of classic signal detection parameters measuring sensitivity and criterion placement. In contrast, the studies I have presented here are subjective – neither the participant nor the researcher can say on a given trial whether a particular color was blue or purple, a face was threatening or nonthreatening, or a study proposal was ethical or unethical. Because participant accuracy cannot be measured objectively, traditional signal detection parameters are ill-suited to capture how participant responses change when the prevalence changes. The lack of ground truth also may mean that on a conscious level, accuracy may be less paramount as a concern for participants in my tasks -- I don't provide them with any feedback or certainty about their performance, either during the task or afterwards.

Forced responses. As Thomson and colleagues write, "the hallmark of laboratory vigilance tasks is that there are very few critical events requiring some sort of response" (2015). Observers in vigilance tasks typically only respond when there is a signal to identify, rather than being forced to provide a response on every trial. Vigilance tasks involve long periods of sitting

and watching for a particular stimulus. Contrast this with the two-alternative forced choice tasks in my studies, in which observers have to identify on every trial whether the stimulus at hand belongs to one concept or another. Thomson et al. speculate that the criterion shift observed in their studies represents a form of mind wandering or inattention. Perhaps in my studies, requiring participants to actively classify every stimulus results in fewer inattentive errors than in traditional vigilance tasks.

*Task difficulty. T*hough my tasks are subjective, I suspect that they are simply easier than the visual search tasks used in studies of the LPE, even in the absence of quantifiable accuracy measures. Identifying a single stimulus on a computer screen is less daunting than a rich task that simulates a real world visual search, as many LPE studies strive to do. If the LPE is a failure of visual search brought on by the complex nature of visual search and the difficulty of looking for rare targets, then my comparatively low-stakes identification tasks may not be taxing enough for such errors to play a large role.

Discrete vs continuous stimuli. Perhaps the biggest difference to note here is the nature of the stimuli that participants are asked to judge. One reason that there is no ground truth in my tasks is that I use stimuli that are drawn from smooth unidimensional continua, rather than discrete kinds. For example, in my studies on color identification, color values increase incrementally from very purple to very blue. In LPE tasks, by contrast, stimuli are often drawn from simple shapes or real-world objects that can be discretely classified as signals or noise. This difference may have an important consequence for how these data can be analyzed. The extreme ends of the spectra used in my studies are usually at floors or ceilings of identification. For example, very blue dots are nearly always called blue, and very purple dots are nearly always

called purple. The spectra for study ethicality and facial animacy exhibit the same property.[6] As a result, the interaction of interest in each study – a three-way interaction of intensity, trial, and prevalence – relies heavily on changing responses for the more ambiguous middle of each spectrum, since participants responses on extreme ends of the spectrum are difficult to change. Intuitively, this makes sense. I am arguing that prevalence-induced concept change is a broadening or contracting of concept boundaries, and those boundaries are usually right at the midpoint of the stimuli spectra used in my studies. However, it raises an important question – do participants in vigilance and LPE studies respond differently to maximally ambiguous stimuli at low prevalence, or are these effects largely insensitive to stimulus intensity? Future work reconciling these phenomena could explicitly try to model such differences by using continuous rather than discrete stimuli continua.

The substantial differences between the tasks used to study vigilance, the LPE, and my findings suggest that these phenomena are not necessarily at odds and might even be found to coexist with the right study design. It is entirely possible that the kinds of tasks in which observers exhibit prevalence-induced concept change could also exhibit both the vigilance decrement and the low prevalence effect. Perhaps, as in the LPE, my participants are proportionally more likely to "miss" a very blue dot or very unethical study when it is displayed at low prevalence. This effect could be present, but masked by the boundary expansion I see, or by the fact that my task is relatively simple compared to many visual search tasks. Regarding vigilance, my average task duration always less than one hour. Perhaps if my tasks went on for

[6] The facial threat stimuli used in Study 7 are an exception. As Figure 9 depicts, participants in this task did not call high-intensity faces threatening 100% of the time. Anecdotally, several participants reported in the post-task debrief that facial threat seemed to be confounded with gender and skin tone in these tasks, and that they did not feel comfortable categorizing the high-intensity faces as more threatening, simply because they appeared to be more male or darker-skinned.

long enough, my participants would eventually suffer a vigilance decrement and make more errors or misidentify more stimuli after observing for long periods of time. The vigilance decrement, low prevalence effect, and prevalence-induced concept change are three different quirks of detection, and given the differences in the circumstances that seem to elicit them, they may even co-occur with the right set of circumstances.

Debiasing strategies

Several of the studies presented here attempted to prevent participants from shifting their decision thresholds in response to prevalence changes. In Study 3, I explicitly warned participants about the impending change in prevalence. In Study 4, I appealed to participants to remain consistent in their responses, and offered financial incentives if they could do so. In Study 5, I shifted the prevalence very quickly, in the hopes that participants would notice the change and compensate in their responses accordingly. In Study 6, I put a visual reference onscreen during the task representing each participant's maximally ambiguous color. In Study 9, I paired color stimuli with facial animacy stimuli that changed in prevalence hydraulically with one another. Thought the presence of a visual reference may have slightly reduced the size of the threshold shift, none of these interventions prevented participants from shifting their decision thresholds when the prevalence changed.

The possibility that prevalence-induced concept change is due to a Range-Frequency process makes the failure of the interventions easier to understand. If prevalence determines how the brain computes the value or intensity of stimuli at a relatively early stage of processing, then interventions that rely on top-down modulation or reappraisal (Studies 3, 4, and 5) may have little ability to change stimuli comparisons, especially if absolute relations between stimuli have already been discarded (or perhaps were not stored to begin with). Study 6, which presented

participants on each trial with a second, maximally ambiguous dot (usually from the middle of the color spectrum) as a visual reference, would do little to change either the range or frequency parameters in the RFT model.[7] Study 9, which involved compensatory changes in animacy prevalence along with color prevalence shifts, also did little to change either the range or frequency of stimuli, and as a result its inability to prevent conceptual shifting is not surprising.

Can the knowledge that Range-Frequency Theory plays a role in prevalence-induced concept change inform new interventions to help prevent it? The clearest strategies 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 parameter may be more feasible. For example, when the prevalence of blue dots is decreased, extending the range of intensity of purple dots (such that participants are exposed to extremely purple dots that they had not seen earlier) would decrease the range of recent stimuli even as the frequency increases. The high-frequency bursts that have been used to compensate for the Low Prevalence Effect in airport baggage screeners (Wolfe et al., 2013) present a model to emulate when designing such interventions.

Limitations and future directions

Prevalence-induced concept change, at least as tested in the studies presented here, seems to be a stubborn phenomenon that occurs in domains as different as color identification and

[7] If, as the results tentatively suggest, the presence of a visual reference reduced the size of the prevalence shift in Study 6, why did this occur? One possibility is that the visual reference, appearing onscreen alongside each target stimulus on each trial in the task, becomes privileged in the observer's temporal context compared to other recently seen stimuli. A modified Range-Frequency model that incorporates weighting based on the recency of items in the contextual set could explore this possibility further.

moral judgment. However, many questions still remain about the scope and duration of this effect.

How many tasks would show the same shift in concept boundaries I have demonstrated with colors, facial threat and research ethics? I suspect that observers are less susceptible to a prevalence-induced shift in domains where instances of different concepts are very easy to distinguish, or have a firm and salient boundary. For example, the visual search tasks used in some research on the low prevalence effect, such as looking for an upside-down letter in an array of letters, present little difficulty in distinguishing a normal letter from an upside-down letter given enough time and attention. Rather, the difficulty is in finding the upside-down letter in the first place. Similarly, tasks where the boundaries between two concepts are specific, firm, and immovable may be less susceptible to this effect. For example, imagine a judge making decisions about granting bail. If the sole criterion for being disqualified from parole was whether or not the crime involved bodily injury, then this would be an unambiguous way to sort crimes – a crime either involves bodily injury, or does not. By contrast, imagine that the criteria instead was "serious injury." This is a more subjective concept, akin to whether or not a face is threatening or a study is unethical. I would predict that prevalence-induced concept change would be more likely, and easier to measure, in the latter case.

Another question involves the role of time in the task, and how long it persists after the intervention. All of my studies involved a participant in the laboratory making judgments in sequence for under an hour. If there were fewer, more intermittent judgments, or if they were more spaced out in time, would the phenomenon persist? On a computational level, the relevant question seems to be what is incorporated into the observer's local context. Study 9 suggests that participants probably partition and monitor different contextual sets for different classes of

judgment (in that case, animacy vs color). Likewise, if an observer is occasionally judging colors, while also doing other tasks in-between and being exposed to a variety of colors in their environment, to what degree can they maintain and track prevalence in the local context of just colors?

Second, once participants are done making judgments, how persistent is the impact of prevalence, and does it ever extend to other judgments? For example, imagine that participants in Study 1 had been shown a picture of the visible light spectrum right after the task, and asked to draw a line representing the boundary between blue and purple on the wheel. Now imagine that a second set of participants did the same thing, but first waited an hour, or a full day. Would the prevalence shift during the laboratory task have any impact of a different kind of elicitation of the conceptual boundary in question? In the case of colors, I suspect that it would have little or no impact, because the extreme values present in a color wheel would present a new, salient range manipulation that would override any range present in memory. However, I suspect that the impact of prevalence manipulations might be more persistent in domains such as Study 8's ethics judgments. Unlike memory of colors, memory of specific ethical violations may be more salient and less subject to interference from new violations. As a result, extremely low- or high-prevalence manipulations in less perceptual domains might have longer lasting impact.

Finally, while all of my attempts to encourage participants to use conscious control to hold conceptual boundaries stable in the face of prevalence change were unsuccessful, one interesting strategy that I did not test here involves alerting participants to the distinction between prevalence-dependent and -independent judgments. In the present studies, this distinction was either not mentioned, or it was implied that these were prevalence-independent judgments. However, a direct manipulation in which participants did the same task, but some

were convinced that it was prevalence-dependent while others were convinced that it was prevalence independent, would serve as a more direct test of the possibility that top-down control can influence judgments, if not perceptions, under conditions of low or high prevalence. This kind of debiasing strategy, if effective, would be particularly useful in dynamic systems in which the observer is responsible for any prevalence shift that occurs. As an example, a fisherman who only wants "big fish" may over time deplete the stock of a lake by catching all of the fish he considers "big." In response to that prevalence shift, my results would suggest that he would then relax his decision threshold for large fish, enabling him to keep catching fish long after all of the fish he initially considered big are gone. Perhaps efforts to remind him that he was responsible for the prevalence shift would alert him to his changed standards. This example relates to existing lines of research about how aware decision-makers are of the circumstances that are creating their decision environments. Most broadly, people tend to neglect situational factors when evaluating the actions and mental states of others (Gilbert & Malone, 1995), as well as the systems responsible for generating the events they experience (Massey & Wu, 2005). More directly relevant to prevalence-independent decisions is the frequent case that observers in those domains are partly responsible for prevalence changes they experience (Gilbert & Jones, 1986), and thus, should know not to adjust their standards as a result.

Many of my conclusions about the meaning of my findings and possible intervention strategies rest on the assumption, supported by the computational models presented here, that Range-Frequency Theory is the best candidate mechanism to describe how humans implement prevalence-induced concept change on a cognitive level. However, I should note that I have only explored a small subset of the possible model space. Many sophisticated models of contextual perception and valuation exist in the cognitive sciences. In Decision by Sampling (Stewart,

Chater, & Brown, 2006), no explicit value of stimuli is directly computed, and valuations come solely from comparisons with local context. It is possible that such a model, or a revised version of one of the models I tested, would better predict my data. Bhui and Gershman (2018) have recently argued that Decision by Sampling and Range-Frequency Theory are in fact related examples of the principle of efficient coding in valuation. This convergence of existing theories of contextual effects on judgment only lends credence to the likelihood that prevalence-induced concept change is, at heart, driven by some implementation of efficient coding in the brain.

Concluding Remarks

Across nine studies, prevalence-induced concept change occurred when it should not have. When blue dots became rare, purple dots began to look blue; when threatening faces became rare, neutral faces began to look threatening; and when unethical research proposals became rare, ambiguous research proposals began to seem unethical. This happened even when the change in the prevalence of instances was abrupt, even when participants were explicitly told that the prevalence of instances would change, and even when participants were instructed and paid to ignore these changes.

The fact that absolute 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 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. Our studies suggest that 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" in the middle of the game. This phenomenon is not limited to IRBs, of course, and seems likely to plague the well-meaning enforcers of any social policy that requires people to ameliorate the very thing they are attempting to assess. The irony is that those who devote themselves to reducing the prevalence of social problems—from unethical proposals to unwarranted aggressions—may not recognize the success of their efforts simply because they view each new instance in the decreasingly problematic context that they themselves have brought about.

A complementary consequence is that observers may quickly become desensitized to problems when they proliferate – exactly the time when accurate classification is most important. 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 leads to more ethical lapses being forgiven. Indeed, research by Gino and Bazerman (2009) has found evidence to this effect when studying cheating behavior.

Modern societies have made extraordinary progress in solving a wide range of social problems, from poverty and illiteracy to violence and infant mortality (Pinker, 2011), and yet, the majority of people believe the world is getting worse instead of better (Dinic, 2016). The fact that concepts grow larger when their instances grow smaller may be a potent source of this pessimism.

REFERENCES

Anstis, S., Verstraten, F. a J., & Mather, G. (1998). The motion aftereffect. *Trends in Cognitive Sciences*, *2*(3), 111–117.

Bartlett, N. R. (1965). *Dark adaptation and light adaptation*. New York: Wiley.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bhargava, S., & Fisman, R. (2014). Contrast effects in sequential decisions: Evidence from speed dating. *The Review of Economics and Statstics*, *96*(3), 444–457.

Bhui, R., & Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *BioRxiv*.

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.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, (November), 1–12.

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.

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., DeGardelle, V., Herce Castanon, S., & Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron*, *81*(6), 1429–1441.

Chen, D. L., Moskowitz, T. J., & Shue, K. (2016). Decision making under the Gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *Quarterly Journal of Economics*, *131*(3), 1181–1242.

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., & Schwartz, O. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research*, *47*(25), 3125–3131.

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied*, *12*(3), 166–178.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and Performance XXIII*, *23*, 3–38.

Decarlo, L. T. (2013). Signal detection models for the same – different task. *Journal of Mathematical Psychology*, *57*(1–2), 43–51.

Dinic, M. (2016). Is the world getting better or worse? Retrieved April 4, 2018, from https://yougov.co.uk/news/2016/01/08/fsafasf/

Epley, N., & Gilovich, T. (2006). The Anchoring-and-Adjustment Heuristic: Why the Adjustments Are Insufficient. *Psychological Science*, *17*(4), 311–318.

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening. *PloS One*.

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.

Fornaciai, M., & Park, J. (2018). Attractive Serial Dependence in the Absence of an Explicit Task. *Psychological Science*, *29*(3), 437–446.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6.

Gilbert, D. T., & Jones, E. E. (1986). Perceiver-induced constraint: Interpretations of self-generated reality. *Journal of Personality and Social Psychology*, *50*(2), 269–280.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*.

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.

Godwin, H. J., Menneer, T., Riggs, C. a., Cave, K. R., & Donnelly, N. (2014). Perceptual failures in the selection and identification of low-prevalence targets in relative prevalence visual search. *Attention, Perception, & Psychophysics*, *77*(1), 150–159.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585–612.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2007). *Bayesian models of cognition*.

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*.

Haslam, N. (2016). Concept creep: psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, *27*(1), 1–17.

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as Reference Points. *Cognitive Psychology*, *109*, 79–109.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2), 181–197.

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

Hout, M. C., Walenchok, S. C., Goldinger, S. D., & Wolfe, J. M. (2015). Failures of Perception in the Low-Prevalence Effect : Evidence From Active and Passive Visual Search. *Journal of*

Experimental Psychology: Human Perception and Performance, *41*(4), 977–994.

Hsu, S.-M., & Young, A. (2004). Adaptation effects in facial expression recognition. *Visual Cognition*, *11*(7), 871–899.

Kahneman, D., Miller, D. T., Griffin, D., Mcpherson, L., & Read, D. (1986). Norm Theory : Comparing Reality to Its Alternatives. *Psychological Review*, *93*(2), 136–153.

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.

Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modelling in r for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, *7*(5), 573–579.

Leopold, D. A., Rhodes, G., Muller, K. M., & Jeffery, L. (2005). The dynamics of visual adaptation to faces. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1566), 897–904.

Lilienfeld, S. O. (2017). Microaggressions: Strong Claims, Inadequate Evidence. *Perspectives on Psychological Science*, *12*(1), 138–169.

Looser, C. E., & Wheatley, T. (2010). The Tipping Point of Animacy: How, When, and Where We Perceive Life in a Face. *Psychological Science*, *21*(12), 1854–1862.

Louie, K., & Glimcher, P. W. (2012). Efficient coding and the neural representation of value. *Ann. N.Y. Acad. Sci*, *1251*(1), 13–32.

Louie, K., Glimcher, P. W., & Webb, R. (2015). Adaptive neural coding: From biological to behavioral decision-making. *Current Opinion in Behavioral Sciences*, *5*, 91–99.

Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward value-based gain control: divisive normalization in parietal cortex. *Journal of Neuroscience*, *31*(29), 10627–10639.

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.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Psychology Press. Retrieved from http://books.google.com/books?id=P094AgAAQBAJ&pg=PA418&dq=inauthor:creelman+signal+detection&hl=&cd=1&source=gbs_api

Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory & Cognition*, *22*(6), 657–672.

Massey, C., & Wu, G. (2005). Detecting Regime Shifts: The Causes of Under- and Overreaction. *Management Science*, *51*(6), 932–947.

Mather, G., Pavan, A., Campana, G., & Casco, C. (2008). The motion aftereffect reloaded. *Trends in Cognitive Sciences*, *12*(12), 481–487.

Matthews, G., & Davies, D. R. (2001). Individual differences in energetic arousal and sustained attention: a dual task study. *Personality and Individual Differences, 31, 575-589 .*, *31*, 575–589.

Mellers, B. A., & Birnbaum, M. H. (1983). Contextual effects in social judgment. *Journal of Experimental Social Psychology*, *19*(2), 157–171.

Mitroff, S. R., & Biggs, A. T. (2013). The Ultra-Rare-Item Effect: Visual Search for Exceedingly Rare Items Is Highly Susceptible to Error. *Psychological Science*.

Nakagawa, S., O'Hara, R. B., & Schielzeth, H. (2012). A general and simple method for

obtaining R2from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

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.

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

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*(6), 407–418.

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.

Peltier, C., & Becker, M. W. (2016). Decision Processes in Visual Search as a Function of Target Prevalence. *Journal of Experimental Psychology: Human Perception and Performance*.

Pinker, S. (2011). *The Better Angels of Our Nature. The better angels of our nature: Why violence has declined*. New York: Penguin.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281.

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.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for

group studies - Revisited. *NeuroImage*, *84*, 971–985.

Schneider, S. (2016). The effects of surrounding positive and negative experiences on risk taking. *Society for Judgment and Decision Making*, *11*(5), 424–440.

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review*, *108*(1), 257–272.

Schwark, J., Sandry, J., & Dolgov, I. (2013). Evidence for a positive relationship between working-memory capacity and detection of low-prevalence targets in visual search. *Perception*, *42*(1), 112–114.

Schwark, J., Sandry, J., MacDonald, J., & Dolgov, I. (2012). False feedback increases detection of low-prevalence targets in visual search. *Attention, Perception, & Psychophysics*, *74*(8), 1583–1589.

Schwartz, O., Hsu, A., & Dayan, P. (2007). Space and time in visual context. *Nature Reviews. Neuroscience*, *8*(7), 522–535.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 1–12.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., & Hirsch, J. (2015). Predictive Perception Codes for Forthcoming in the Frontal Cortex. *Science*, *314*(5803), 1311–1314.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*, 409.

Temple, J. G., Warm, J. S., Dember, W. N., Jones, K. S., LaGrange, C. M., & Matthews, G. (2000). The Effects of Signal Salience and Caffeine on Performance, Workload, and Stress in an Abbreviated Vigilance Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *42*(2), 183–194.

Thompson, P., & Burr, D. (2009). Visual aftereffects. *Current Biology*, *19*(1), 11–14.

Thomson, D. R., Besner, D., & Smilek, D. (2015). A Critical Examination of the Evidence for Sensitivity Loss in Modern Vigilance Tasks. *Psychological Review*, *123*(1), 1–15.

Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*(4), 724–738.

Todorov, A., & Oosterhof, N. (2011). Modeling Social Perception of Faces [Social Sciences]. *IEEE Signal Processing Magazine*, *28*(2), 117–122.

Trimmer, P. C., Ehlman, S. M., McNamara, J. M., & Sih, A. (2017). The erroneous signals of detection theory. *Proceedings of the Royal Society. Series B, Biological Sciences*, *284*(20171852).

Tversky, A., & Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dpendent Model. *Quarterly Journal of Economics*, *106*(4), 1039–1061.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science*, *39*(10), 1179–1189.

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.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance Requires Hard Mental Work and Is Stressful. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 433–441.

Webster, M. (1996). Human colour perception and its adaptation. *Network: Computation in Neural Systems*, *7*(4), 587–634.

Webster, M. A. (2015). Visual adaptation. *Annual Review of Vision Science*, *1*(1), 547–569.

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.

Wickens, T. D. (2001). *Elementary Signal Detection Theory*. Oxford University Press.

Wiggs, C. L., & Martin, A. (1998). Properties and mechanics of perceptual priming. *Current Opinion in Neurobiology*, *8*(1), 227–233.

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.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*, *136*(4), 623–638.

Wolfe, J. M., & Van Wert, M. J. (2010). Varying Target Prevalence Reveals Two Dissociable Decision Criteria in Visual Search. *Current Biology*, *20*(2), 121–124.

APPENDIX I: INSTRUCTIONS FOR ALL STUDIES

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 20 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.
[*after each series:*
Series complete.
Please take a short break. We'll start the next series in a moment.]

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 20 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.
[*after each series:*
Series complete.
Please take a short break. We'll start the next series in a moment.]

Study 3 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 20 series.
[*Stable prevalence with warning condition only:*
As the study goes on, blue dots are not going to become more or less common. In other words, you will see the same amount of them over time. Your only task is to identify blue dots.]
[*Decreasing prevalence with warning condition only:*
As the study goes on, blue dots are going to become less common. In other words, you will see fewer of them over time. 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.
[*after each series:*
Series complete.
Please take a short break. We’ll start the next series in a moment.]

Study 4 Instructions

Welcome to this study! We are 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 others may have only a few.
[*decreasing + instruction* and *decreasing + instruction + incentive* conditions only:
Some of the dots you see will appear more than once. Please try to be consistent throughout the study. Once you have identified a dot as blue or not blue, please do your best to respond the same way if you see that dot again later.]
[*decreasing + instruction + incentive* condition only:
As an incentive, we will be awarding a bonus of $10 to the five most consistent participants in this study.]
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, don't worry -- just keep going.
Now you will complete a brief practice series so you can see how the task works.
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.
[*after each series:*
Series complete.

Please take a short break. We'll start the next series in a moment.]

Study 5 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.
[*after each series:*
Series complete.
Please take a short break. We'll start the next series in a moment.]

Study 6 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, don't worry – 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.
[*after each series:*
Series complete.
Please take a short break. We'll start the next series in a moment.]
[after the first 200 trials]
At this point in the study, the task is going to change. Please press the spacebar to continue and read about these changes.
So far, you have been deciding if each dot you saw was "blue" or "not blue."

Based on your responses in the task so far, we have calculated your personal "blue standard." Your standard represents the least blue dot that you still called blue. In other words, it's your definition of where the color blue begins.
Now, there will be two dots on the screen. On the left, you will see your standard. This dot won't change.
On the right, you will see the new dot that you have to judge. The buttons work the same way as they did before.
Remember, you're only judging the dot on the right side of the screen. If you have any questions, you should ask the experimenter now. Otherwise, press the space bar to continue the study.

Study 7 Instructions

Welcome to this study! We're interested in studying how people perceive and identify faces.
In this task, you will see faces presented on the screen one at a time. Your task in this study will be to identify faces with threatening facial
expressions.
When you see a threatening face on the screen, press the "threat" key. For all other faces, press the "no threat" key.
The faces will be presented in series with breaks in between. This means that you will see a series of faces, have a short break, and then another series of faces, until you have seen 16 series.
Some of the series you see may have a lot of threatening faces, and other may have only a few. There's nothing for you to count or keep track of -- your only task is to identify threatening faces.
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.
[*after each series:*
Series complete.
Please take a short break. We'll start the next series in a moment.]

Study 8 Instructions

Welcome to this study! We're interested in studying how people make ethical decisions about scientific experiments.
Many scientific experiments involve some risk for the participants because they can cause psychological distress or physical harm. Universities have to make difficult ethical decisions about whether or not to allow experiments to be conducted.
Today, you will read about various experiments that could be conducted on human beings. We simply want to know whether you think scientists SHOULD or SHOULD NOT be allowed to conduct each of these experiments.
Because this is an ethical decision, there are no right or wrong answers. We simply want your personal decision for each study.
Here are some things to keep in mind as you make your decisions.

1) All of the experiments you will read about will be conducted on adults who have volunteered to take part in exchange for money.
2) All of the experiments are part of research on human behavior.
3) When scientists must lie to the participants either before or during the experiment, they always tell the participants the truth when the experiment is over.
4) Participants are always free to withdraw and can stop participating at any time they wish.
In the task, you will see descriptions of experiments presented on the screen,
one at a time.
When you read a description of an experiment that you would not allow to be conducted, press the "REJECT" key. For all other experiments, press the "APPROVE" key.
The experiments will be presented in series, with breaks in between. This means that you will read a series of experiments, have a short break, and then another series of experiments, until you have seen 10 series.
Some of the series you see may have a lot of unethical experiments, and others may have only a few. There's nothing for you to count or keep track of -- your only task is to approve or reject each experiment.
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 round so you can get used to the task.
You have now completed the practice round. If you have any questions, you can ask the experimenter now. Otherwise, you're ready to begin the study.
[*after each series:*
Series complete.
Please take a short break. We'll start the next series in a moment.]

Study 9 Instructions

Welcome to this study! We are interested in studying how people perceive and identify colors and faces.
In this task, you will see two types of images: dots and faces. The images will be presented on the screen one at a time. Your tasks in this study will be to identify blue dots and animate (living) faces.
When you see a blue dot or an animate face on the screen, press the "blue/animate" key. For all other dots and faces, press the "not blue/inanimate" key.
The images will be presented in series with breaks in between. This means that you will see a series of images, have a short break, and then another series of images, until you have seen 16 series.
Some of the series you see may have a lot of blue dots or animate faces, and others may have only a few.
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, don't worry -- just keep going.
Now you will complete a brief practice series so you can see how the task works.
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.
[*after each series:*
Series complete.

Please take a short break. We'll start the next series in a moment.]

APPENDIX II: POST-TASK QUESTIONNAIRES FOR ALL STUDIES

Studies 1 and 2 Post-task questionnaire

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

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

How old are you?

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)

Do you have normal color vision? If not, please be specific about the nature of your color blindness (ex: red-green colorblind, blue-yellow colorblind, etc).
- ❍ Yes, I have normal color vision (1)
- ❍ No (explain) (2) _____

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)

[if no] Since you indicated that you grew up speaking languages other than English, please tell us what language(s) you grew up speaking: _____

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 sliders 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)

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

_____

Study 3 Post-task questionnaire

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

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

How old are you?

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)

Do you have normal color vision? If not, please be specific about the nature of your color blindness (ex: red-green colorblind, blue-yellow colorblind, etc).

- ❍ Yes, I have normal color vision (1)
- ❍ No (explain) (2) _____

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)

[if no] Since you indicated that you grew up speaking languages other than English, please tell us what language(s) you grew up speaking:

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 sliders 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 defintiion 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)

What did the instructions at the beginning of the study say about how the amount of blue dots would change over time?

- ❍ The instructions said that there would be fewer blue dots over time (1)
- ❍ The instructions said that there would be more blue dots over time (2)
- ❍ The instructions said that the amount of blue dots over time would not change (3)
- ❍ I'm not sure (4)

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

Study 4 Post-task questionnaire

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

Please indicate your gender:

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

How old are you?

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)

Do you have normal color vision? If not, please be specific about the nature of your color blindness (ex: red-green colorblind, blue-yellow colorblind, etc).

- ❍ Yes, I have normal color vision (1)
- ❍ No (explain) (2) _____

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)

[if no] Since you indicated that you grew up speaking languages other than English, please tell us what language(s) you grew up speaking:

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 sliders 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 defintiion 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)

What did the instructions at the beginning of the study say about the consistency of your responses in the task?

- ❍ The instructions said that I should try to be consistent over time (1)
- ❍ The instructions said that I shouldn't worry about trying to be consistent over time (2)
- ❍ The instructions didn't mention consistency at all (3)
- ❍ I'm not sure (4)

Were you told that it was possible to earn a cash bonus in this experiment?

- ❍ Yes, a bonus awarded randomly (1)
- ❍ Yes, a bonus for the most consistent participants (2)
- ❍ No, a cash bonus was not mentioned at all (3)
- ❍ I'm not sure (4)

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

Study 5 Post-task questionnaire

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

Please indicate your gender:

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

How old are you?

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)

Do you have normal color vision? If not, please be specific about the nature of your color blindness (ex: red-green colorblind, blue-yellow colorblind, etc).

- ❍ Yes, I have normal color vision (1)
- ❍ No (explain) (2) _____

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)

[if no] Since you indicated that you grew up speaking languages other than English, please tell us what language(s) you grew up speaking:

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 sliders 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 defintiion 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)

What did the instructions at the beginning of the study say about how the amount of blue dots would change over time?

- ❍ The instructions said that there would be fewer blue dots over time (1)
- ❍ The instructions said that there would be more blue dots over time (2)
- ❍ The instructions said that the amount of blue dots over time would not change (3)
- ❍ I'm not sure (4)

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

Study 6 Post-task questionnaire

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

Please indicate your gender:

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

How old are you?

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)

Do you have normal color vision? If not, please be specific about the nature of your color blindness (ex: red-green colorblind, blue-yellow colorblind, etc).
- ❍ Yes, I have normal color vision (1)
- ❍ No (explain) (2) _____

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)

[if no] Since you indicated that you grew up speaking languages other than English, please tell us what language(s) you grew up speaking:

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 sliders 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 defintiion 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 7 Post-task questionnaire

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

Please indicate your gender:

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

How old are you?

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)

Do you have normal color vision? If not, please be specific about the nature of your color blindness (ex: red-green colorblind, blue-yellow colorblind, etc).

- ❍ Yes, I have normal color vision (1)
- ❍ No (explain) (2) _____

Do you have trouble recognizing faces, or do you have prosopagnosia (face-blindness)?

- ❍ No, neither (1)
- ❍ Yes, I just have trouble recognizing faces (2)
- ❍ Yes, I have prosopagnosia (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)

[if no] Since you indicated that you grew up speaking languages other than English, please tell us what language(s) you grew up speaking:

What do you think this study was about?

Do you think it became easier or harder to find threatening faces as the study progressed?

- ❍ It became easier to find threatening faces as the study progressed (1)
- ❍ It became harder to find threatening faces 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 threatening faces in each series changed during the study?

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

We want to get a sense of how many threatening faces you think you saw at different times in the study. Please indicate, using the sliders below, your impressions about what proportion of the faces you saw were threatening. 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)

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

Study 8 Post-task questionnaire

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

Please indicate your gender:

- ❍ Male (1)
- ❍ Female (2)
- ❍ Other: (3) _____
- ❍ Prefer not to answer (4)

How old are you?

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)

[if no] Since you indicated that you grew up speaking languages other than English, please tell us what language(s) you grew up speaking:

Do you think that it became easier or harder to find unethical studies as the study progressed?
- ❍ It became easier to find unethical studies as the study progressed (1)
- ❍ It became harder to find unethical studies 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 unethical experiments in each series changed during the study?
- ❍ No, there were the same number of unethical experiments throughout the study (1)
- ❍ Yes, there were fewer unethical experiments as the study went on (2)
- ❍ Yes, there were more unethical experiments as the study went on (3)
- ❍ I'm not sure (4)

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

the experiments you saw were unethical. 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 an "unethical" experiment changed?

- ❍ No, I think that my definition of what counted as an unethical experiment did not change during the study. (1)
- ❍ Yes, I think my definition of what counts as an unethical experiment expanded -- I counted a wider range of experiments as unethical at the end of the study compared to the beginning of the study. (2)
- ❍ Yes, I think my definition of what counts as an unethical experiment narrowed -- I counted a smaller range of experiments as unethical 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)

What do you think this study was about?

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

Study 9 Post-task questionnaire

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

Please indicate your gender:

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

How old are you?

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)

Do you have normal color vision? If not, please be specific about the nature of your color blindness (ex: red-green colorblind, blue-yellow colorblind, etc).

- ❍ Yes, I have normal color vision (1)
- ❍ No (explain) (2) _____

Do you have trouble recognizing faces, or do you have prosopagnosia (face-blindness)?

- ❍ No, neither (1)
- ❍ Yes, I just have trouble recognizing faces (2)
- ❍ Yes, I have prosopagnosia (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)

[if no] Since you indicated that you grew up speaking languages other than English, please tell us what language(s) you grew up speaking:

What do you think this study was about?

Do you think 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 think it became easier or harder to find animate faces as the study progressed?
❍ It became easier to find animate faces as the study progressed (1)
❍ It became harder to find animate faces 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)

Do you feel that the amount of animate faces in each series changed during the study?
❍ No, there were the same number of animate faces throughout the study (1)
❍ Yes, there were fewer animate faces as the study went on (2)
❍ Yes, there were more animate faces 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 sliders 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)

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

By the end of the study, do you think that your definition of what counted as an "animate" face changed?

- ❍ No, I think that my definition of what counted as an animate face did not change during the study. (1)
- ❍ Yes, I think my definition of what counts as an animate face expanded -- I counted a wider range of faces as animate at the end of the study compared to the beginning of the study. (2)
- ❍ Yes, I think my definition of what counts as an animate face narrowed -- I counted a smaller range of faces as animate 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)

What does it mean for a face to be “animate?”

- ❍ That it looks alive
- ❍ That it doesn’t look alive
- ❍ I don’t know

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