The central tendency bias in color perception: Effects of internal and external noise

Maria Olkkonen

Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania

Patrice F. McCarthy

Department of Psychology, Rutgers, The State University of New Jersey, Camden, NJ

Sarah R. Allred

Department of Psychology, Rutgers, The State University of New Jersey, Camden, NJ

Perceptual estimates can be biased by previously seen stimuli in delayed estimation tasks. These biases are often toward the mean of the whole stimulus set. Recently, we demonstrated such a central tendency bias in delayed color estimation. In the Bayesian framework of perceptual inference, perceptual biases arise when noisy sensory measurements are combined with prior information about the world. Here, we investigate this idea in color perception by manipulating stimulus range and stimulus noise while characterizing delayed color estimates. First, we manipulated the experimental prior for stimulus color by embedding stimuli in collections with different hue ranges. Stimulus range affected hue bias: Hue estimates were always biased toward the mean of the current set. Next, we studied the effect of internal and external noise on the amount of hue bias. Internal noise was manipulated by increasing the delay between the reference and test from 0.4 to 4 s. External noise was manipulated by increasing the amount of chromatic noise in the reference stimulus, while keeping the delay between the reference and test constant at 2 s. Both noise manipulations had a reliable effect on the strength of the central tendency bias. Furthermore, there was a tendency for a positive relationship between variability of the estimates and bias in both noise conditions. In conclusion, observers are able to learn an experimental hue prior, and the weight on the prior can be manipulated by introducing noise in the estimation process.

Introduction

The visual appearance of objects is determined not only by the immediate context in which they are viewed, but also by previous experience, both on a timescale of an experiment (Adams, Graf, & Ernst, 2004; Ashourian & Loewenstein, 2011; Chopin & Mamassian, 2012; Jazayeri & Shadlen, 2010; Olkkonen & Allred, 2014), as well as on a timescale of an individual's lifetime or longer (Blake & Bülthoff, 1990; Girshick, Landy, & Simoncelli, 2011; Hansen, Olkkonen, Walter, & Gegenfurtner, 2006; Kleffner & Ramachandran, 1992; Konkle & Oliva, 2007; Stocker & Simoncelli, 2006; Welchman, Lam, & Bülthoff, 2008; Witzel, Valkova, Hansen, & Gegenfurtner, 2011). We recently reported a hue appearance bias in a delayed estimation task for a center-surround display, along with a strong interaction between the delay bias and an appearance shift caused by spatial color contrast (Olkkonen & Allred, 2014). Delayed hue estimates were biased toward the mean hue of all displayed stimuli, indicating that observers used information extracted from the stimulus set to estimate memorized hue. Similar central tendency biases have been reported for estimates of size (Hollingworth, 1910), shape and gray value (Huttenlocher, Hedges, & Vevea, 2000), line length (Ashourian & Loewenstein, 2011; Duffy, Huttenlocher, Hedges, & Crawford, 2010; Huttenlocher et al., 2000) and interval duration (Jazayeri & Shadlen, 2010). Previous studies on color memory have not found consistent evidence for hue memory biases, although it has often been suggested that hue memory might be biased toward focal colors (for discussion see e.g., Ling & Hurlbert, 2008). The first aim of this paper is to verify the existence of a central tendency bias for hue by measuring the effect of stimulus range on delayed hue estimates.

We (Olkkonen & Allred, 2014) and others (e.g., Ashourian & Loewenstein, 2011; Girshick et al., 2011) have also reported a relationship between variability

and bias in perceptual estimates: More variable estimates tend to be more biased. This kind of relationship between variability and bias can be modeled with Bayesian approaches to perceptual estimation (e.g., Ashourian & Loewenstein, 2011; Girshick et al., 2011; Jazayeri & Shadlen, 2010; Stocker & Simoncelli, 2006; Weiss, Simoncelli, & Adelson, 2002). For example, Ashourian and Loewenstein (2011) showed that a Bayesian model accounted for a central tendency bias in a delayed line length estimation task. In their model, the first and second line stimuli were both modeled as noisy representations. However, since the first line was retained in memory, it was represented with more variability than the second line. Hence, prior information had more influence on the estimate for the first line, causing an overall bias toward the mean length of the whole set. Ashourian and Loewenstein (2011) also manipulated internal noise by introducing a color working memory task during the delay period. Adding noise increased both variability and response errors. Importantly, and as predicted by the Bayesian model, the increase in variability was related to the amount of central tendency bias. We hypothesized that a similar mechanism could account for the hue biases in our previous report (Olkkonen & Allred, 2014). Although we did not manipulate variability in that study, variability correlated with the absolute magnitude of bias across data sets.

If observers indeed build a prior based on the stimulus set in a delayed hue comparison task, manipulating the variability of the hue representation should affect the strength of hue bias. Although the short-term retention of perceptual information is surprisingly robust, there is some increase in variability over short retention intervals for many stimulus dimensions (for a review see Pasternak & Greenlee, 2005). For color, the increase in variability seems to be fast at very short delays, stabilizing after a few seconds (Nemes, Parry, & McKeefry, 2010; Nilsson & Nelson, 1981). However, whether changes in color appearance accompany these changes in variability is more contested; biases in saturation and lightness for delayed estimates have consistently been found (de Fez, Capilla, Luque, Pérez-Carpinell, & del Pozo, 2001; Hanawalt & Post, 1942; Newhall, Burnham, & Clark, 1957), but this is not the case for hue (e.g., de Fez et al., 2001; Ling & Hurlbert, 2008; Nilsson & Nelson, 1981). The second aim of this paper is to systematically measure the effect of internal noise on both appearance and variability of hue estimates, where internal noise is manipulated by varying the length of the retention interval.

If biases in hue perception indeed arise because noisy sensory representations are combined with prior information, then any increase in sensory variability should increase bias. Girshick et al. (2011) showed that adding orientation noise to a stimulus increased

orientation discrimination thresholds and led to more bias toward cardinal orientations. In the color domain, external noise in the form of increased chromatic variability leads to higher color discrimination thresholds for simultaneously presented stimuli (Giesel, Hansen, & Gegenfurtner, 2009; Hansen, Giesel, & Gegenfurtner, 2008; Li & Lennie, 1997; te Pas & Koenderink, 2004), but it is unknown what effects external noise has on color appearance, and whether there is a relationship between thresholds and bias for an external noise manipulation. The final aim of this paper is to investigate whether external noise as defined by chromatic variability in the stimulus also elicits concomitant changes in hue variability and appearance.

To summarize, here we investigate the extent to which stimulus collection and internal and external noise affect the central tendency bias in delayed hue estimation. First, we manipulated the experimental hue prior by varying the range of stimuli between experimental runs. Second, we increased internal noise, operationalized by delay between reference and test stimulus. Third, we increased external noise, operationalized by chromatic variability in the reference stimulus.

Experiment 1

The purpose of Experiment 1 was to establish the source of the central tendency bias in delayed hue estimation. If at least some part of the hue bias is experiment-specific, then the bias for a given hue in an experiment should be affected by the overall range of hues presented during that experiment. If, on the other hand, the hue bias arises from a longer-term prior, such as bias toward focal green, then the range of hues presented during an experiment should have no effect on bias. To that end, delayed hue estimates were collected in three blocks with different but overlapping hue ranges.

Methods

Observers

Eight observers naive to the purpose of the study participated. Observers had normal or corrected to normal visual acuity and normal color vision as assessed with the Ishihara color plates. Observers were undergraduate students at Rutgers University who signed informed consent and received either pay ($10/hour) or course credit for their time. The research protocol was approved by the university Institutional Review Board.
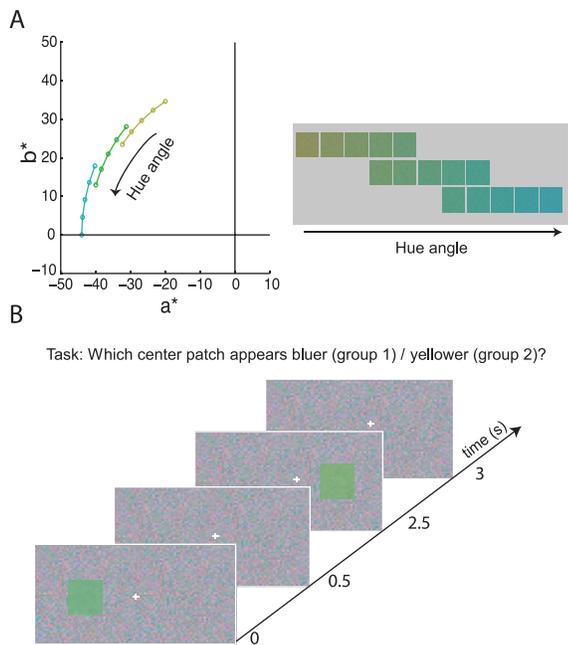
Figure 1. Stimuli and task in Experiment 1. (A) Reference chromaticities in the a*b* plane of the CIELAB space for three stimulus ranges. All stimuli had the same saturation; the range radii are offset in the graph for clarity. Approximate renderings of the stimuli are shown on the right. See Table 1 for color specifications. (B) Stimulus timing. On each trial, the reference was shown for 0.5 s. After a 2 s interstimulus interval, the test was shown for 0.5 s. Reference hue range was varied between blocks. Trials for each reference hue within a block were interleaved. The observers' task on each trial was to select the stimulus that appeared bluer (group 1) or yellower (group 2).

Stimuli

The reference and test stimuli were 1.8°/visual angle squares displayed 3° on the left or on the right of a central fixation cross. Eleven equally spaced reference stimuli were selected from a hue circle on an equiluminant plane of the CIELAB space. Three stimulus ranges were

employed in three blocks of trials. Each range had two stimuli that overlapped with the adjacent range (see Figure 1A and Table 1 for stimulus color specifications). The ranges were selected by eye to correspond to bluish-green, green, and yellowish-green hues.

Test stimulus hue was controlled with an adaptive staircase procedure (see Procedure). Saturation (CIE-LAB radius 40) and luminance (15 cd/m^2) of the references and tests were fixed.

Both the reference and test stimuli as well as the background consisted of a checkerboard texture (check size 0.2° × 0.2°). The luminance of the stimulus checks was perturbed around the mean display value of 15 cd/m^2. The checks in the background were perturbed in both luminance and chromaticity around the mean xyY values of the background (0.313 0.392 15).

Apparatus

Stimuli were displayed on a calibrated CRT monitor (1024 × 786 pixels/24° × 18°; 85 Hz) with a 10-bit intensity resolution per color channel via the Datapixx box (VPixx Technologies, Inc., Saint-Bruno-de-Montarville, Quebec, Canada). The monitor was calibrated once a month with standard methods (Brainard, Pelli, & Robson, 2002).

MGL functions (URL: http://gru.brain.riken.jp/doku.php/mgl/) were used for stimulus display and data collection in Matlab (Mathworks, Inc., Natick, MA).

Procedure

Observers viewed the display from a 94 cm distance controlled with a chin rest. On each trial, the reference was displayed for 500 ms; after a 2 s interstimulus interval, the test was displayed on the opposite side of the display for 500 ms (Figure 1). After the offset of the test, the observer indicated on a button box which stimulus appeared bluer or yellower, depending on task assignment (see Methods/Task). The response initiated

Reference #	Set	Hue angle	x	y	Y cd/m^2	a*	b*	L*
1	1	120	0.342	0.407	15	−20.0	34.6	100
2	1	126	0.334	0.406	15	−23.5	32.4	100
3	1	132	0.325	0.404	15	−26.8	29.7	100
4	1, 2	138	0.317	0.400	15	−29.7	26.8	100
5	1, 2	144	0.308	0.396	15	−32.4	23.5	100
6	2	150	0.299	0.391	15	−34.6	20.0	100
7	2, 3	156	0.291	0.385	15	−36.5	16.3	100
8	2, 3	162	0.283	0.378	15	−38.0	12.4	100
9	3	168	0.275	0.370	15	−39.1	8.3	100
10	3	174	0.268	0.362	15	−39.8	4.2	100
11	3	180	0.262	0.354	15	−40.0	0	100

Table 1. CIE 1931 xyY and CIE L*a*b* values of the reference stimuli for each set. Note that some references belong to two sets. Y and L* values correspond to the mean of the normal distribution from which the luminances were sampled for each stimulus check.

the next trial. The left/right locations of the reference and test on each trial were randomized.

Test hue for each reference was controlled with a staircase procedure. Two staircases tracked roughly the 30th and 70th percentiles of each psychometric function, with starting points below and above the reference hue, respectively. Each staircase had 15 trials. The two staircases for the five reference stimuli in a given set were interleaved in one block, resulting in 150 trials per block.

The three different stimulus ranges were employed in different blocks. Block order was counterbalanced across observers. Observers S3 and S7 ran each block twice; the other observers ran each block once. Raw data for S3 and S7 were pooled across repetitions before data analysis. Each session with one or more blocks took between 45–60 minutes.

Task

When observers are asked to evaluate a stimulus on a continuum between two endpoints, e.g., blue-yellow, they may show a response bias that depends on the question they are being asked: "Which stimulus is bluer?" vs. "Which stimulus is yellower?" (see Jogan & Stocker, 2014, for discussion). Indeed, in addition to a central tendency bias, we found a response bias in our previous study, in which we asked observers which of two stimuli was bluer (Olkkonen & Allred, 2014). The response bias was a uniform shift toward yellow, i.e., away from the indicated color word. Thus, here we split our subject group into two: One half participated in the "yellow" task, and the other half in the "blue" task. We assumed that averaging the data across the two tasks circumvented any response bias solely due to the polarity of the task.

Data analysis

From the bluer (yellower) responses, the probability was calculated at each test hue of selecting the test stimulus as bluer (yellower) than the reference. Psychometric functions (PMFs) were estimated by fitting cumulative normals to the proportion-test-responses data with the Psignifit package (Wichmann & Hill, 2001a, 2001b). Color appearance of a reference was defined as the 50th percentile of the PMF. This point denotes the test hue which is perceptually indistinguishable from the reference hue (point of subjective equality, PSE). Bias was defined as the difference between a given PSE and veridical reference hue. To quantify variability, we defined hue discrimination thresholds as the hue angle spanned by the difference between the 75th and 50th percentile of the PMFs.

The central tendency bias was characterized by fitting a linear regression model with slope and offset to

the bias as a function of reference hue for each of the three ranges. We assumed that a central tendency bias would be manifested by a negative slope as a function of reference. Furthermore, for any negative slopes, the size of the slope would quantify the strength of the bias.

Results

Bias and variability estimates were derived from psychometric functions, as shown in Figure 2 for one observer. The two psychometric functions in Figure 2A show data for the same reference hue in two different range conditions. The true reference hue is indicated by the dashed vertical line, and the points of subjective equality (PSE) for each psychometric function are shown with colored vertical lines. If stimulus range had no effect on responses, the two psychometric functions should overlap; however, they are clearly distinct. We defined bias as the hue angle between the true reference hue and each PSE. For this reference hue, the stimulus range clearly affected PSE: For the range shown in grey, the bias is toward smaller color angles ("yellow"), whereas for the other range (red), the estimate is biased in the opposite direction, toward larger color angles ("blue"). Thus, the same physical hue was perceived differently in delayed estimation when it was embedded in two different temporal hue contexts.

In Figure 2B, we plot bias as a function of reference hue for each of the three stimulus ranges. Biases calculated from the psychometric functions in Figure 2A are indicated with circles. Inspection of this figure reveals two salient points. First, the bias for a given reference hue depends on range. Note that reference hues presented in multiple ranges are indicated by gray panels: If stimulus range had no effect on PSE, then these pairs of data points should overlap. Instead, they differ substantially. Second, PSEs in each reference hue range exhibited an overall bias towards the central stimulus in that range. For each range, the vertical line indicates the central stimulus: Reference hues near these lines exhibit little bias, whereas other reference hues are biased toward the central stimulus.

In Figure 2C, we plot discrimination thresholds, defined as the difference between the 75th and 50th percentile of the psychometric function, for the same observer. Although there was substantial variability in thresholds for the different reference hues, we did not observe systematic variability.

The biases averaged over eight observers are shown in Figure 3A. As with observer S3, bias for a given reference hue depended on context: The data points in the grey bands differ consistently. Within each range, the average bias for smaller hue angles was positive, indicating a bias towards larger hue angles, whereas the
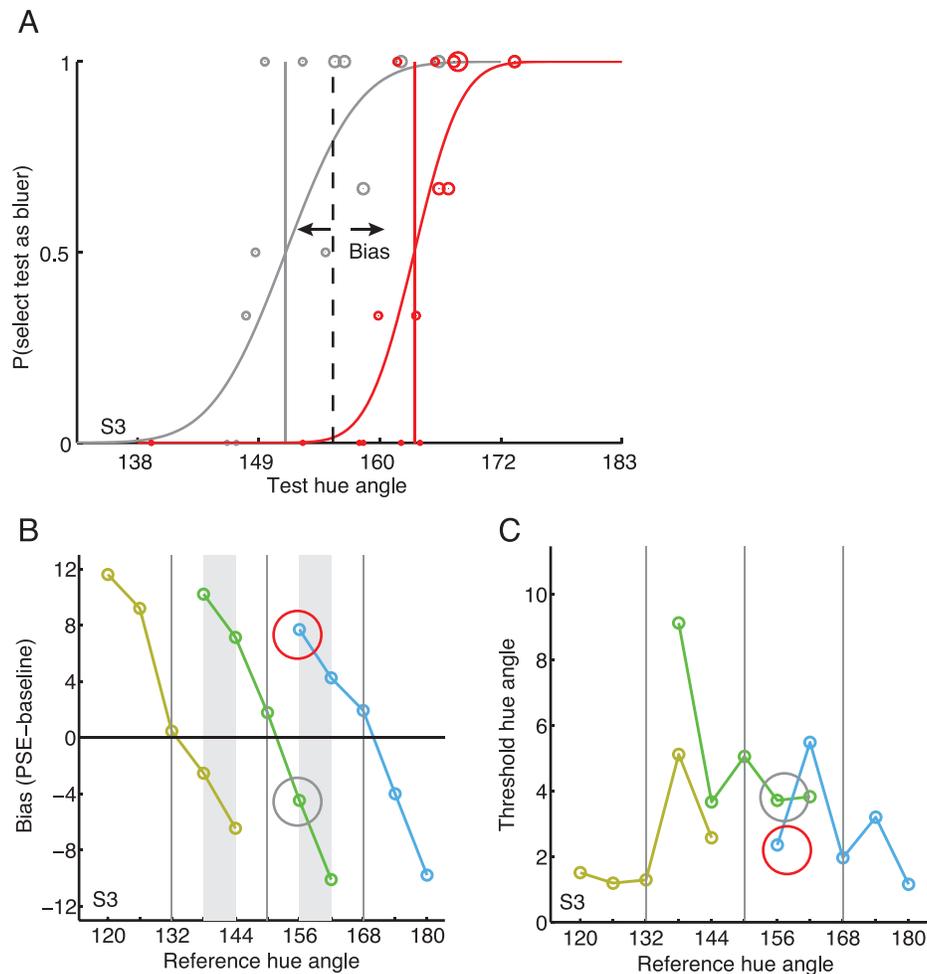
Figure 2. Deriving bias and thresholds from psychometric functions in Experiment 1. (A) Psychometric functions are shown for observer S3 for one reference in two range conditions. Test hue varies on the *x* axis, while probability of selecting the test as bluer is on the *y* axis. Each data point is the proportion of trials on which the test hue indicated on the *x* axis was judged as bluer. Lines are cumulative normal fits to the data. The vertical dashed line indicates the reference hue; solid vertical lines indicate the PSEs for the two functions. (B) Bias, calculated as the difference between the PSE and reference hue, is plotted against reference hue for the three ranges. The gray and red circles indicate the reference hue and stimulus ranges for which the psychometric functions are shown in (A). (C) Discrimination thresholds, defined by the difference between the 70th and 50th percentile of the psychometric function, are shown for each range across reference hue.

average bias for larger hue angles was negative, indicating an overall bias toward smaller hue angles.

We quantified the magnitude of this effect by fitting lines to the bias; the slope of the line indicates the strength of the bias, while the intercept indicates the hue toward which appearance is biased. Furthermore, since the middle hue in each range represents the mean hue of that range, then intercepts near the middle hue point to a central tendency bias. To test whether this was the case, we compared the fitted intercepts to intercepts that would produce a line crossing exactly at the middle reference hue. The predicted and fitted intercepts were virtually identical (mean fitted vs. predicted intercepts for the three ranges: 33.2/33.2, 41.9/43.0, 45.3/46.4; all corrected *p* values in *t* tests > 0.3).

Slopes of the bias across reference hue for each observer and range are shown in Figure 3B. Since the intercepts were close to the middle reference hue, the negative slope of the bias is a measure of the strength of the central tendency bias. Six out of eight observers had negative slopes for some or all the range sets. Range did not affect slope in a mixed-effects ANOVA with range as fixed factor and observer as random factor, $F(2, 14) = 0.15$, $p = 0.86$. In separate *t* tests, the slopes for the two rightmost ranges were significantly negative (Bonferroni-corrected one-tailed *p* values 0.12, 0.03, 0.03, for the three ranges).

Although reference hue range affected the PSE for a given hue, the range did not affect thresholds systematically. This result is seen in Figure 3C, where the average thresholds for each range are similar. In
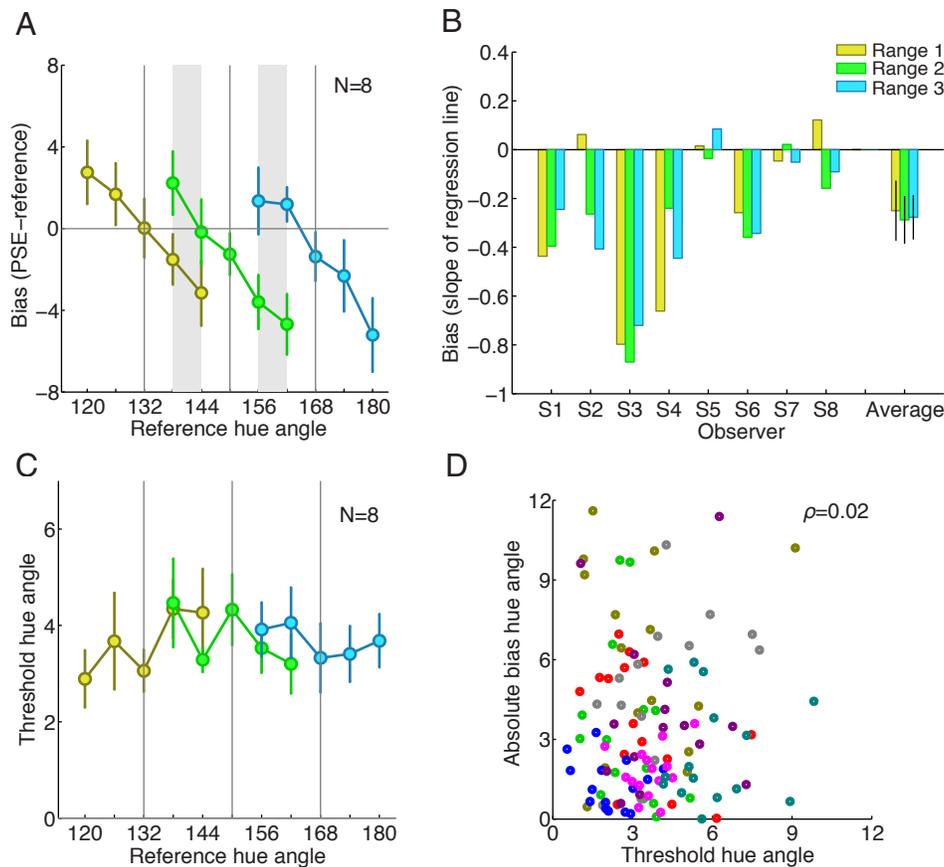
Figure 3. Effect of hue range on bias and variability in Experiment 1. (A) Bias, calculated as the difference between the PSE and veridical reference hue, is plotted against reference hue for the three reference hue ranges. Vertical lines represent the central hue in each reference range. Error bars are ± 1 *SEM* across eight observers. Shaded areas indicate the reference hues common to adjacent ranges. (B) Central tendency bias was quantified as the slope of the bias across reference stimuli. Each set of bars is for one observer; the rightmost set of bars is the average across observers, with error bars indicating ± 1 *SEM*. (C) 75% discrimination thresholds are plotted against reference hue angle for the three ranges. Vertical lines represent the central hue in each reference range. Error bars are ± 1 *SEM*. (D) Absolute value of bias is plotted against discrimination threshold for each reference hue, range, and observer. Symbol color indicates observer. Correlation coefficient is shown on top right.

addition, reference hue angle overall did not seem to affect variability: Thresholds for small reference hue angles did not differ from thresholds for large hue angles, independent of range. This outcome is consistent with a recent report that discrimination thresholds are not affected by category boundaries in the blue/green region (Witzel & Gegenfurtner, 2013).

There was no correlation between absolute bias magnitude and thresholds across observers, range, and reference hue ($\rho = 0.02$), indicating that bias magnitude for individual data sets was unrelated to variability.

central stimulus value in each range. That bias can be manipulated by stimulus range falsifies the common, albeit unverified, assumption that hue biases are caused by internal prototypes brought by observers to the experiment (see e.g., Nilsson & Nelson, 1981). The result is instead consistent with observers developing a hue prior in the course of the experiment, which they combine with a variable measurement of the reference stimulus in order to estimate its hue (Ashourian & Loewenstein, 2011; Duffy et al., 2010; Huttenlocher et al., 2000; Jazayeri & Shadlen, 2010; Olkkonen & Allred, 2014).

Summary

The results of the first experiment show that hue bias in delayed estimation tasks can be manipulated by the range of stimuli presented during the course of an experiment. Furthermore, the bias tends toward the

Experiment 2

Increasing delay between reference and test can affect the variability in color estimates (Nemes et al., 2010; Newhall et al., 1957; Nilsson & Nelson, 1981),

A

B   Task: Which center patch appears bluer (group 1) / yellower (group 2)?
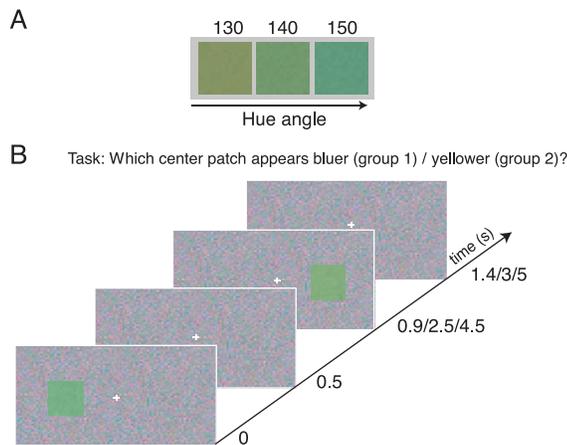
Figure 4. Stimuli and task in Experiment 2. (A) Approximate renditions of the reference stimuli. Hue angle increases from left to right. See Table 2 for color specifications. (B) Stimulus timing. The reference was displayed first for 0.5 s; after a variable delay (0.4, 2, or 4 s), the test was displayed for 0.5 s. Trials for all references and delay lengths were interleaved within a block of trials. The observers' task on each trial was to select the stimulus that appeared bluer (group 1) or yellower (group 2).

Reference #	Hue angle (degree)	x	y	Y (cd/m^2)	a*	b*	L*
1	130	0.328	0.405	15	−25.7	30.6	100
2	140	0.134	0.399	15	−30.6	25.7	100
3	150	0.299	0.391	15	−34.6	20.0	100

Table 2. CIE 1931 xyY and CIE L*a*b* values of the reference stimuli. Y and L* values correspond to the mean of the normal distribution from which the luminances were sampled for each stimulus check.

reference hues were sampled from the blue-green part of the CIELAB space (see Figure 4A and Table 2 for color values). The saturation (CIELAB radius 40) and luminance (15 cd/m^2) of the stimuli were fixed. Timing was otherwise as in Experiment 1, but the delay between the reference and test on a given trial was either 0.4, 2 or 4 s (Figure 4B). Trials for the different delays were interleaved.

Five of the observers ran the "Which stimulus appears bluer" version of the task, while four observers ran the "Which stimulus appears yellower" task.

A staircase procedure was used to control the test hue. Four interleaved staircases tracked roughly the 20th, 30th, 70th, and 80th percentiles of the psychometric function with the first two and last two staircases having starting points on opposite sides of the reference hue. With three reference stimuli, three delays, four staircases, and 20 trials per staircase, there were 720 trials in total per observer. The experiment was divided into three parts, which were run on different days. Three observers (S9, S13, and S14) ran the entire experiment twice, while the rest ran it once.

Psychometric functions for each reference in each condition were estimated by fitting cumulative normals to the test-selected-as-bluer (yellower) responses with the Psignifit package. For the observers who ran the experiment twice, the raw responses were pooled across repetitions before fitting the PMFs. PSEs, bias and variability were derived from the fits as in Experiment 1.

but little is known about how delay affects bias. We hypothesized that increasing the delay between reference and test would increase internal noise and cause observers to rely more on prior information. From this hypothesis, we predicted that increasing delays would increase both thresholds and the magnitude of the central tendency bias. The purpose of Experiment 2 was to test these predictions. To this end, three different delays were selected for the interval between reference and test, and we measured thresholds and bias for three reference hues for each delay.

Methods

Observers

Observer S1 and nine new observers, all naive to the purpose of the study, participated under the same subject protocol as in Experiment 1. Observers had normal or corrected-to-normal visual acuity and normal color vision as assessed with the Ishihara color plates. One observer was excluded before data analysis because of nonconverging staircases, leaving nine observers.

Stimuli and procedure

The apparatus was the same as in Experiment 1; the monitor was calibrated once a month with standard methods. Geometric display properties, such as stimulus size and location, were as in Experiment 1. Three

Results

A central tendency bias emerged with increasing delay between reference and test. Figure 5A shows hue bias averaged across observers for each delay as a function of reference hue. For the shortest delay (red circles), there was a small "yellow" bias for the reference with the smallest ("yellow") hue angle, causing an overall bias with positive slope across reference hue. For the longer delays, the bias changed sign, indicating an emerging central tendency bias: The reference with the smallest hue angle exhibited a bias toward larger hue angles, and the reference with the
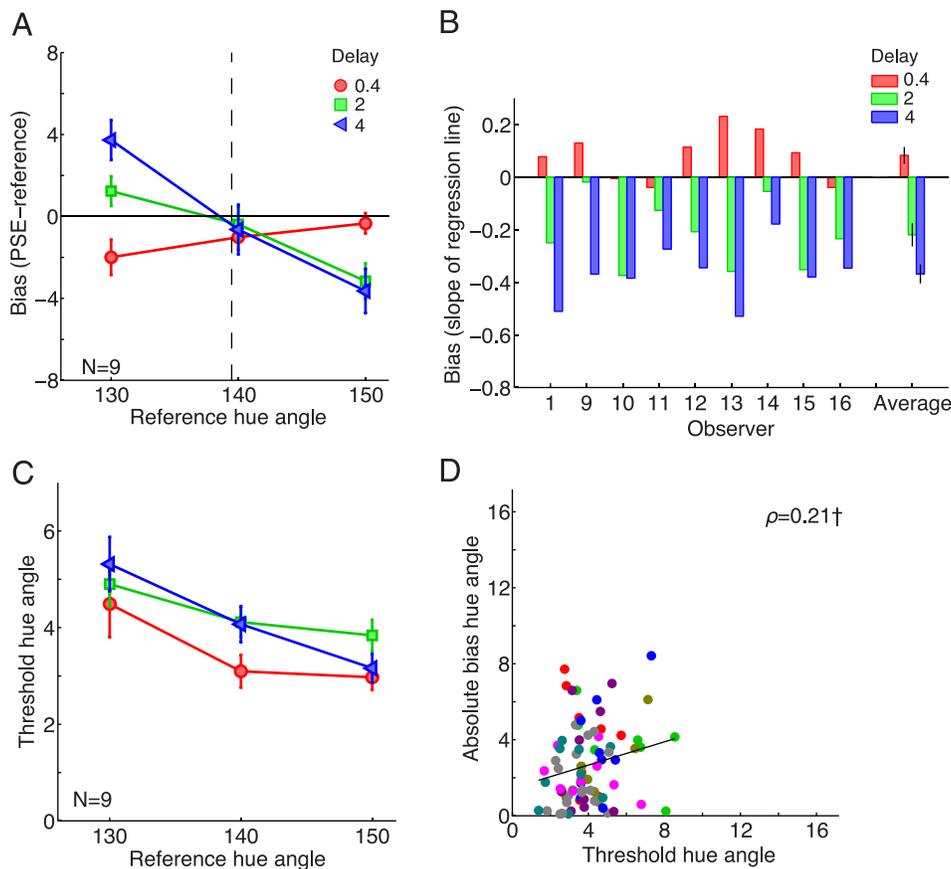
Figure 5. Effect of delay length on bias and thresholds in Experiment 2. (A) Hue bias, defined as the difference between the PSE and veridical reference hue, is plotted against reference hue for each of the three delays. Error bars are ± 1 *SEM* across nine observers. The horizontal solid line indicates zero bias. The vertical dashed line indicates the mean hue of all displayed stimuli (references and tests), averaged across observers. (B). Each bar represents the slope of the line that best fit bias as a function of reference hue for a given delay. Each set of bars is for one observer; the rightmost set of bars shows the average slopes across observers with error bars indicating ± 1 *SEM*. (C) 75% discrimination thresholds in degrees are plotted for each delay as a function of reference hue. Error bars are ± 1 *SEM*. (D) Absolute value of bias is plotted against threshold for each observer, reference, and delay. Different observers are indicated by symbol colors. Correlation coefficient is shown on top right ($p = 0.053$).

largest hue angle exhibited a bias toward smaller hue angles.

To examine the central tendency effect within observers in more detail, we found best-fit lines for bias as a function of reference hue. As in Experiment 1, we compared the fitted intercepts to intercepts predicted from a perfect central tendency bias; the differences were not significant (all corrected *p* values > 0.2). Thus, we used the slope of the lines to quantify the strength of the central tendency bias. The individual slopes, along with the average, are shown in Figure 5B. For the longer delays, the slopes became increasingly negative for all observers, indicating a strengthening central tendency bias. The main effect of the delay manipulation on absolute bias magnitude was significant in a three-way ANOVA with reference and delay as fixed factors and observer as random factor, $F(2, 32) = 10.6$, $p = 0.001$. The effect of reference was not significant, nor was the interaction of reference with delay.

In contrast to the effect of delay on bias, we found only a weak effect of delay on thresholds. In Figure 5C, thresholds are plotted as a function of reference hue for each delay. The effect of delay on thresholds was only marginally significant, $F(2, 32) = 3.21$, $p = 0.07$; post-hoc multiple comparisons showed that the difference between the shortest and medium delay reached significance ($p < 0.05$, Bonferroni-corrected). Reference hue a had larger effect on thresholds: Thresholds decreased toward larger ("bluer") hue angles, $F(2, 32) = 13.7$, $p < 0.001$.

Despite the weak overall effect of delay length on thresholds, we hypothesized that there might be an indirect relationship mediated by observers' individual levels of bias. The relationship between thresholds and bias across all conditions and observers, shown in Figure 5D, offers some support for this hypothesis; there was a marginally significant correlation of 0.21 between thresholds and bias ($p = 0.053$).
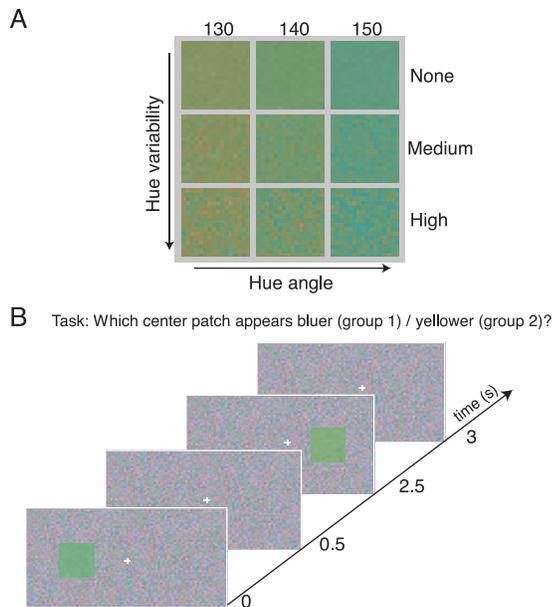
Figure 6. Stimuli and task in Experiment 3. (A) Approximate renditions of the reference stimuli. Hue angle increases from left to right; chromatic noise (variability of individual check hue around mean hue angle) increases from top to bottom. All stimuli had the same amount of luminance noise. B) Stimulus timing. The reference was displayed first for 0.5 s; after a 2 s ISI, the test was displayed for 0.5 s. Trials for all references and noise levels were interleaved within a block of trials. The observers' task on each trial was to select the stimulus that appeared bluer (group 1) or yellower (group 2).

Summary

Hue estimates at the shortest delay were slightly positively biased. For longer delays, the bias reversed sign and strengthened, indicating a central tendency bias. There was also a significant increase in thresholds from the shortest to the medium delay, but no further increase to the longest delay. Finally, there was a trend for a positive relationship between thresholds and bias across observers.

Experiment 3

Color discrimination with nonuniformly colored stimuli has rarely been characterized, but based on the few existing reports, discrimination performance tends to deteriorate with increasing chromatic noise (Giesel et al., 2009; Hansen et al., 2008; Li & Lennie, 1997; te Pas & Koenderink, 2004). We hypothesized that increasing external noise in the form of chromatic variability in the reference would cause observers to rely more on prior information when making delayed hue estimates. Fol-

lowing this hypothesis, we predicted that increasing chromatic noise would increase both thresholds and the magnitude of the central tendency bias observed. To test these predictions, we selected one delay (2 s) and three different levels of chromatic noise, and measured bias and thresholds at each level of noise for three reference hues.

Methods

Observers

S1 along with eight new observers, all naive to the purpose of the study, participated under the same subject protocol as in Experiments 1 and 2. Observers had normal or corrected-to-normal visual acuity and normal color vision as assessed with the Ishihara color plates.

Stimuli and procedure

Stimuli were displayed on the same apparatus as in Experiments 1 and 2, calibrated once a month with standard methods. Geometric display properties were as in Experiment 1. The three reference hues from Experiment 2 were employed (Table 2). The chromatic noise of the reference stimuli was manipulated by sampling the hue for the stimulus checks in each reference from a hue distribution with zero, medium, or high variability ($SD = 0°$, $8.5°$, and $17°$ in the a*b* place of CIELAB space, respectively; Figure 6A). The test stimulus always had zero chromatic noise. Stimulus timing was identical to Experiment 1 (Figure 6B).

A group of four observers ran the "Which stimulus appears bluer" task, while another group of five observers ran the "Which stimulus appears yellower" task.

A two-stage procedure was employed to characterize psychometric functions (Olkkonen & Allred, 2014). Psychometric functions were measured with a staircase procedure in the first part of the experiment. Two interleaved staircases tracked roughly the 30th and 70th percentiles of each psychometric function, with respective starting points below and above the reference hue. Each staircase had 15 trials. Staircases for the three different reference stimuli and three noise levels were interleaved, resulting in 270 trials per block. Cumulative normals were fitted to the proportion-test-selected-as-bluer (yellower) data in each condition. Based on these fits, the test hues that corresponded roughly to 0% and 100% probability to select the test as bluer (yellower) were chosen as endpoints for the next phase of the experiment. In this phase, data were collected for five test levels evenly distributed between the endpoints with the method of constant stimuli (MOCS). Specifically, on each trial the test hue paired with a given reference was randomly selected from the five predetermined values. Each test level was repeated
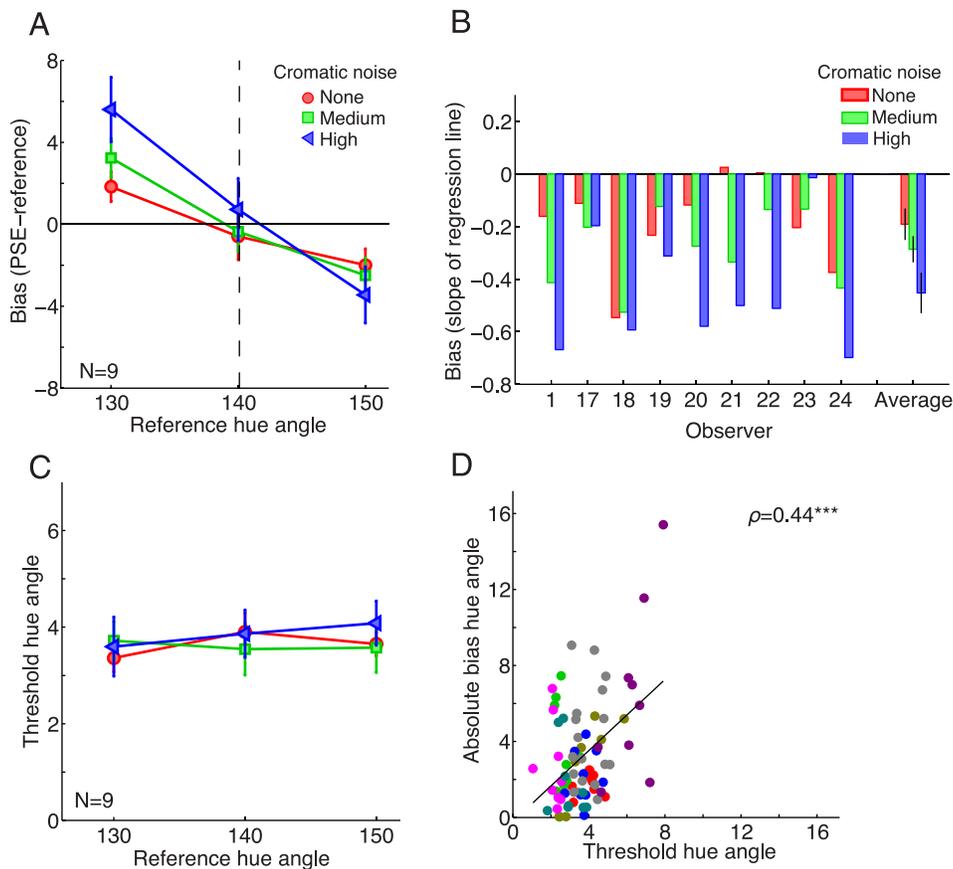
Figure 7. Effect of chromatic noise on bias and thresholds in Experiment 3. (A) Hue bias, defined as the difference between the PSE and veridical reference hue, is plotted against reference hue for each of the three chromatic noise levels. Error bars are ±1 SEM across nine observers. The horizontal solid line indicates zero bias. The vertical dashed line indicates the mean hue of all displayed stimuli (references and tests), averaged across observers. (B) Each bar represents the slope of the line that best fit bias as a function of reference hue for a given noise level. Each set of bars is for one observer; the rightmost set of bars is the average across observers, with error bars indicating ± 1 SEM. (C) 75% discrimination thresholds are plotted for each noise level as a function of reference hue. Error bars are ±1 SEM. (D) Absolute value of bias is plotted against threshold for each observer, reference, and noise level. Different observers are indicated with symbol colors. Correlation coefficient is shown on top right ($p < 0.001$).

10 times during one block, with the three references and three noise levels interleaved. After finishing one complete MOCS run, the test hue range was adjusted as needed, after which the MOCS experiment was repeated.

For data analysis, the staircase and MOCS data were pooled. Psychometric functions were estimated and bias and thresholds quantified as in Experiments 1 and 2.

Results

Hue bias was sensitive to the amount of chromatic noise in the reference stimulus, confirming our first prediction. Figure 7A shows the average bias for each noise level and reference hue. As in Experiments 1 and 2, responses to hue angles smaller than the reference stimulus exhibited a bias toward larger

angles, and response to hue angles larger than the reference stimulus exhibited a bias toward smaller angles. Furthermore, this effect was exaggerated with increasing chromatic noise, as indicated by the blue points (higher chromatic noise) falling farther from the zero bias line than the red points (low chromatic noise). A three-way ANOVA with reference and noise as fixed factors and observer as random factor confirmed a main effect of the noise manipulation on absolute bias magnitude, $F(2, 32) = 7.6$, $p = 0.005$.

To examine the central tendency effect within observers in more detail, we again found best-fit lines for bias as a function of reference hue. The fitted intercepts were not significantly different from intercepts predicted from a perfect central tendency bias (all corrected p values > 0.26). The central tendency effect, quantified by the slope across reference hue as in Experiments 1 and 2, was smallest for zero chromatic
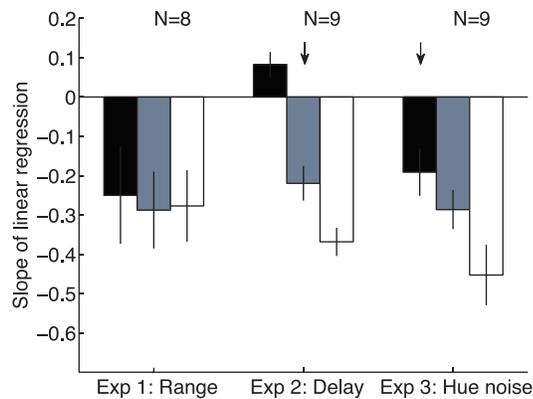
Figure 8. Comparison of central tendency bias across experiments. Each set of bars shows average slopes in each experiment with error bars ($\pm$ 1 *SEM*). In Experiment 1, there were three overlapping hue ranges. In Experiment 2, there were three delays (0.2, 2, 4 s). In Experiment 3, there were three levels of chromatic noise (zero, medium, high). Identical conditions are marked with arrows.

noise, and became larger with increasing chromatic noise. This was the case for all but one observer (S23; Figure 7B).

In contrast to the clear effect of chromatic noise on the central tendency bias, the chromatic noise manipulation did not have a significant effect on thresholds, as shown in Figure 7C, three-way mixed-effects ANOVA, $F(2, 32) = 0.63$, $p = 0.54$. However, there was again substantial interindividual variability in thresholds, and we hypothesized that although there was no average effect of chromatic noise on thresholds, there might be an indirect relationship mediated by observers' individual levels of bias. Indeed, thresholds and bias across observers, reference, and noise level were significantly correlated ($\rho = 0.46$, $p < 0.001$, Figure 7D). Thus, data sets with higher thresholds tended to deviate more from zero bias. To further test the effect of the chromatic noise on bias, we ran a linear regression with absolute bias magnitude as regressand and thresholds, noise level, observer, and reference hue as regressors. In the best-fit model ($r^2 = 0.28$), thresholds accounted for 21% of the variance in the bias, while noise level accounted for 7%, $F(2, 78) = 15.1$, $p < 0.001$. Observer or reference hue did not contribute to the variance significantly, indicating that the relationship between bias and thresholds was not driven solely by a few extreme data points.

Summary

Consistent with our first prediction, adding chromatic noise to the reference stimulus increased the magnitude of the central tendency bias. The support from these data for the second prediction that chromatic noise would increase thresholds is less clear: Although thresholds did not on average increase with chromatic noise, thresholds and bias were moderately correlated across observers.

Comparison between experiments

Figure 8 shows the average slopes from the best-fit lines to the bias for all experiments and conditions. The 2-s delay condition in Experiment 2 is comparable to the 2 s, zero noise condition in Experiment 3 (bars marked with arrows), and indeed, the regression slopes are very similar. All conditions in Experiment 1 are more or less comparable to these, because they employed a 2-s delay and zero chromatic noise. The average slopes are of similar magnitude to the 2-s delay slopes in Experiments 2 and 3.

Discussion

Here we have demonstrated a robust hue bias in delayed estimation, whose strength can be manipulated with the introduction of internal or external noise. Moreover, by manipulating the stimulus range, the bias could be shifted on the hue axis. The hue bias measured here tended toward the central stimulus value rather than toward a single prototype. This observation is consistent with observers learning the hue statistics of the stimuli, and employing those statistics in the estimation of the reference hue.

Previously, we suggested that the hue bias is caused by combining prior information with a noisy sensory measurement (Olkkonen & Allred, 2014), similarly to the line length and time interval estimation models by Ashourian and Loewenstein (2011) and Jazayeri and Shadlen (2010), respectively. Indeed, this formulation led us to hypothesize that adding either internal or external noise to the color representation would decrease the reliability of the sensory information, leading observers to rely more on priors. We predicted that decreased reliability of sensory information would be observable as increased thresholds, and that greater reliance on priors would be observable as increases in the central tendency bias. Although the effects of both noise manipulations on hue bias were robust, the effects on variability were less clear. In the remainder of the paper we discuss the bias and threshold findings in light of the previous literature.

Memory biases for color

In our hands, hue was almost always biased toward the central value of the stimulus collection. Hue biases either toward a central value or toward a prototype have not been consistently reported; this is in contrast to saturation and lightness, for which relatively consistent biases have been found (de Fez et al., 2001; Hanawalt & Post, 1942; Newhall et al., 1957). What can account for this discrepancy? We speculate that the absence of hue bias in previous studies results from the stimulus collections employed: Most if not all color memory studies have employed collections spanning a large portion of the hue circle (see e.g., Nemes et al., 2010; Nilsson & Nelson, 1981). The mean hue of such a collection is close to neutral, thus leading to no measurable central tendency bias. Indeed, central tendency biases are consistently found in other stimulus domains presumably because stimuli such as line length, size, and weight vary on one-dimensional continua and therefore have a well-defined central tendency. This is also true for saturation and lightness (for which a central tendency bias was reported by Huttenlocher et al., 2000).

Models of central tendency bias

The general phenomenon of delayed stimulus estimates being drawn toward a central value is well characterized in the literature (e.g., Ashourian & Loewenstein, 2011; Duffy et al., 2010; Hollingworth, 1910; Huttenlocher et al., 2000; Jazayeri & Shadlen, 2010), and several models have been proposed to account for the effect. In Helson's adaptation level theory (Helson, 1948), central tendency biases can be regarded as attractive biases toward the adaptation-level, which is the neutral point of the system as determined by the weighted mean of the stimulus set. The theory is useful as a general framework for contextual effects in vision, but it does not specify the relationship between noise and bias, which is the main topic here. In the more recent category adjustment model (CAM), stimulus estimates are drawn toward category centers when category-level information (e.g., small/large, light/dark) is combined with a noisy fine-grained memory representation (Duffy et al., 2010; Huttenlocher et al., 2000). Noisier representations are more strongly biased toward category centers. The CAM is conceptually similar to the Bayesian models by Jazayeri and Shadlen (2010) and Ashourian and Loewenstein (2011), which account for central tendency biases as arising from combining a noisy memory representation with a stimulus prior derived from the stimulus collection. Even though in the present study the stimuli were selected such that they were not easy to categorize, it is conceivable that the whole stimulus set formed one category toward whose center the estimates were drawn. In this sense, our results are consistent with the CAM as well as the models by Ashourian and Loewenstein (2011) and Jazayeri and Shadlen (2010).

Relationship between thresholds and bias

Although the introduction of both internal and external noise had the predicted effect on bias, we found no systematic effect of chromatic noise on thresholds, and only a marginally significant effect of delay on thresholds. There are several potential explanations. First, it may be that hue bias and threshold measurements are functionally related as we hypothesized, but that our measurements were insufficiently sensitive to detect this relationship. It seems conceivable that estimating the slope of a psychometric function requires more data than estimating its mean; thus, the PSE measurements might have been more stable than the threshold measurements. This idea is supported by noting that threshold measurements in all experiments were variable between observers. Furthermore, in Experiment 3 there was a significant correlation between thresholds and bias when analyzed across all data points for observers and reference stimuli. In other words, an individual dataset with a high threshold was likely to exhibit a larger bias as well. In Experiment 2, we observed a similar, albeit weaker, correlation between bias and thresholds across all data points. With respect to this correlation, we note that we previously demonstrated a much stronger correlation in a similar experiment (Olkkonen & Allred, 2014). Thus, it is possible that our threshold measurements were not sensitive enough in this particular case to detect the effect of internal noise.

A second alternative explanation is that hue bias is related to reliability of the sensory signal, but that discrimination thresholds do not fully reflect sensory reliability. There are some indications in previous literature that this might be the case: Nemes et al. (2010) and Nilsson and Nelson (1981) found rapid increases in successive color discrimination thresholds at delays roughly between 0 s and 2 s, after which thresholds more or less plateaued. Nemes et al. (2010) also found an increase in hue appearance errors (bias) from the shortest to the longest (10 s) delay. In other words, even after thresholds plateaued, hue appearance continued to shift. This is consistent with our findings in Experiment 2: We observed that delay caused a nonsaturating increase in bias, while thresholds plateaued after the medium delay. It is thus possible that decreased sensory reliability caused by the delay might not be robustly reflected in threshold measurements.

Finally, it may be that bias is unrelated to the reliability of the sensory signal, so that biases and threshold effects in memory occur independently. This suggestion is inconsistent with the Bayesian framework that has been successfully implemented to account for appearance biases in short-term memory for other stimulus domains (Jazayeri & Shadlen, 2010, Ashourian & Loewenstein, 2011). Although the relationship between bias and thresholds in memory has not been studied extensively, perceptual studies of simultaneous contextual color processing (Heinemann, 1961; Hillis & Brainard, 2005, 2007a, 2007b ; Rinner & Gegenfurtner, 2000) provide relevant data. The results are mixed: some contextual changes elicit appearance and threshold changes that are best explained by a common mechanism (Heinemann, 1961; Hillis & Brainard, 2005, 2007b), while other contextual changes do not (Hillis & Brainard, 2007a; Rinner & Gegenfurtner, 2000).

Given that the data on the relationship between bias and thresholds are inconclusive both in color memory and color perception, the issue clearly warrants more study. Resolving this issue conclusively will require more measurements of thresholds and appearance in conditions that elicit memory biases. Here we have demonstrated a family of such conditions. Because simultaneously characterizing thresholds and appearance requires large numbers of trials, however, future measurements along these lines should focus on a smaller subset of conditions.

Conclusions

Our results establish that hue bias in a delayed estimation task is shifted by the stimulus collection in an experiment. Thus, this hue bias is unlikely to be caused by a long-term prior based on color prototypes. In addition, increasing either internal or external noise caused a systematic and reliable increase in the strength of the hue bias. We hypothesized that this increased bias would result from an increase in the variability of sensory information caused by the introduction of noise. Although thresholds on average were not much affected by the noise manipulation, there was a tendency toward a positive relationship between variability and bias across data sets in both noise conditions. We conclude that bias is indeed manipulated in predictable ways by the introduction of noise. We also conclude that more research, perhaps including a different operational definition for the variability of sensory information, is needed to establish the underlying mechanisms responsible for the effect of noise on hue bias.

Keywords: color appearance, color discrimination, memory, probabilistic inference

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the "light-from-above" prior. *Nature Neuroscience, 7*(10), 1057–1058.

Ashourian, P., & Loewenstein, Y. (2011). Bayesian inference underlies the contraction bias in delayed comparison tasks. *PLoS ONE, 6*(5), e19551.

Blake, A., & Bülthoff, H. (1990). Does the brain know the physics of specular reflection. *Nature, 343*(6254), 165–168.

Brainard, D. H., Pelli, D. G., and Robson, T. (2002). Display characterization. In J. Hornak (Ed.), *Encyclopedia of imaging science and technology*. (pp. 172–188). Hoboken, NJ: Wiley.

Chopin, A., & Mamassian, P. (2012). Predictive properties of visual adaptation. *Current Biology, 22*(7), 622–626.

de Fez, M. D., Capilla, P., Luque, M. J., Pérez-Carpinell, J., & del Pozo, J. C. (2001). Asymmetric colour matching: Memory matching versus simultaneous matching. *Color Research & Application, 26*(6), 458–468.

Duffy, S., Huttenlocher, J., Hedges, L. V., & Crawford, L. E. (2010). Category effects on stimulus estimation: Shifting and skewed frequency distributions. *Psychonomic Bulletin & Review, 17*(2), 224–230.

Giesel, M., Hansen, T., & Gegenfurtner, K. R. (2009). The discrimination of chromatic textures. *Journal of Vision, 9*(9):11, 1–28, http://www.journalofvision.org/content/9/9/11, doi:10.1167/9.9.11. [PubMed] [Article]

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience, 14*(7), 926–932.

Hanawalt, N. G., & Post, B. E. (1942). Memory trace for color. *Journal of Experimental Psychology, 30*(3), 216–227.

Hansen, T., Giesel, M., & Gegenfurtner, K. R. (2008). Chromatic discrimination of natural objects. *Journal of Vision, 8*(1):2, 1–19, http://www.journalofvision.org/content/8/1/2, doi:10.1167/8.1.2. [PubMed] [Article]

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience, 9*(11), 1367–1368.

Heinemann, E. G. (1961). The relation of apparent brightness to the threshold for differences in luminance. *Journal of Experimental Psychology, 61*, 389–399.

Helson, B. Y. H. (1948). Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological Review, 55*(6), 297–313.

Hillis, J. M., & Brainard, D. H. (2005). Do common mechanisms of adaptation mediate color discrimination and appearance? Uniform backgrounds. *Journal of the Optical Society of America A, 22*(10), 2090–2106.

Hillis, J. M., & Brainard, D. H. (2007a). Distinct mechanisms mediate visual detection and identification. *Current Biology, 17*(19), 1714–1719.

Hillis, J. M., & Brainard, D. H. (2007b). Do common mechanisms of adaptation mediate color discrimination and appearance? Contrast adaptation. *Journal of the Optical Society of America A, 24*(8), 2122–2133.

Hollingworth, H. I. (1910). The central tendency of judgment. *The Journal of Philosophy, 7*(17), 461–469.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129*(2), 220–241.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience, 13*(8), 1020–1026.

Jogan, M., & Stocker, A. A. (2014). A new two-alternative forced choice method for the unbiased characterization of perceptual bias and discriminability. *Journal of Vision, 14*(3):20, 1–18, http://www.journalofvision.org/content/14/3/20, doi:10.1167/14.3.20. [PubMed] [Article]

Kleffner, D. A., & Ramachandran, V. S. (1992). On the perception of shape from shading. *Perception & Psychophysics, 52*(1), 18–36.

Konkle, T., & Oliva, A. (2007). Normative represen-tation of objects: evidence for an ecological bias in object perception and memory. In D. S. McNamara & J. G. Trafton, (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 407–413). Austin, TX: Cognitive Science Society.

Li, A., & Lennie, P. (1997). Mechanisms underlying segmentation of colored textures. *Vision Research, 37*(1), 83–97.

Ling, Y., & Hurlbert, A. (2008). Role of color memory in successive color constancy. *Journal of the Optical Society of America A, 25*(6), 1215–1226.

Nemes, V. A., Parry, N. R. A., & McKeefry, D. J. (2010). A behavioural investigation of human visual short term memory for colour. *Ophthalmic & Physiological Optics, 30*(5), 594–601.

Newhall, S. M., Burnham, R. W., & Clark, J. R. (1957). Comparison of successive with simultaneous color matching. *Journal of the Optical Society of America, 47*(I), 216–227.

Nilsson, T. H., & Nelson, T. M. (1981). Delayed monochromatic hue matches indicate characteristics of visual memory. *Journal of Experimental Psychology: Human Perception and Performance, 7*(1), 141–150.

Olkkonen, M., & Allred, S. R. (2014). Short-term memory affects color perception in context. *PLoS ONE, 9*(1):e8648, 1–11.

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience, 6*(2), 97–107.

Rinner, O., & Gegenfurtner, K. R. (2000). Time course of chromatic adaptation for color appearance and discrimination. *Vision Research, 40*, 1813–1826.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9*(4), 578–585.

te Pas, S. F., & Koenderink, J. J. (2004). Visual discrimination of spectral distributions. *Perception, 33*(12), 1483–1497.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5*(6), 598–604.

Welchman, A. E., Lam, J. M., & Bülthoff, H. H. (2008). Bayesian motion estimation accounts for a surprising bias in 3D vision. *Proceedings of the National Academy of Sciences, USA, 105*(33), 12087–12092.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and

goodness of fit. *Perception & Psychophysics, 63*(8), 1293–1313.

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics, 63*(8), 1314–1329.

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision, 13*(7):1*,* 1–33, http://www.journalofvision.org/content/13/7/1, doi:10.1167/13.7.1. [PubMed] [Article]

Witzel, C., Valkova, H., Hansen, T., & Gegenfurtner, K. R. (2011). Object knowledge modulates colour appearance. *i-Perception, 2,* 13–49.