Probabilistic visual attentional guidance triggers "feature avoidance" response errors

William Narhi-Martinez, Jiageng Chen, and Julie D. Golomb

Department of Psychology, The Ohio State University

Author Note

Correspondence concerning this article should be addressed to William Narhi-Martinez, Department of Psychology, The Ohio State University 1835 Neil Ave, Columbus, OH 43210, United States. Email: narhi-martinez.1@osu.edu

Word count excluding title, references, author affiliations, acknowledgments, figures and figure legends, and abstract: 11875

Abstract

Spatial attention affects not only where we look, but also what we perceive and remember in attended and unattended locations. Previous work has shown that manipulating attention via top-down cues or bottom-up capture leads to characteristic patterns of feature errors. Here we investigated whether experience-driven attentional guidance – and probabilistic attentional guidance more generally – leads to similar feature errors. We conducted a series of pre-registered experiments employing a learned spatial probability or probabilistic pre-cue; all experiments involved reporting the color of one of four simultaneously presented stimuli using a continuous response modality. When the probabilistic cues guided attention to an invalid (nontarget) location, participants were less likely to report the target color, as expected. But strikingly, their errors tended to be clustered around a nontarget color *opposite* the color of the invalidly-cued nontarget. This "feature avoidance" was found for both experience-driven and top-down probabilistic cues, and appears to be the product of a strategic – but possibly subconscious – behavior, occurring when information about the features and/or feature-location bindings outside the focus of attention is limited. The findings emphasize the importance of considering how different types of attentional guidance can exert different effects on feature perception and memory reports.

Keywords: attention, probabilistic cues, feature perception, mixture modeling, working memory

Public Significance Statement

This study revealed that guiding visual attention to where a target is *likely* to appear impacts how the features of that target are processed in a unique way, different from previously reported feature errors induced when attention is rapidly shifted, divided, or captured away by a distractor. The findings highlight the importance of considering not only where attention is focused but how it got there, and also of using tools that allow for a precise reporting of what we perceive, encode, and recall following the allocation of spatial attention.

Introduction

The world has innumerable locations and objects that we could be directing our visual attention towards at any given moment, and our tendency to be constantly shifting our attention means that what we focus on is often also frequently changing. Our ability to accurately remember, and later report, the features of an object depend greatly on where our attention is being allocated. There has been a wealth of literature focusing on how and why we direct our attention to certain locations or objects, and at present these influences of attention are usually assigned to one of three broad influences: top-down, bottom-up, or experience-driven (Awh et al., 2012). Top-down influences of attention originate from the goals of the observer, often elicited by explicit instructions or direct cues. Bottom-up influences are produced by the salience of the stimuli themselves, such as their uniqueness in color, size, shape, etc. Experience-driven attentional guidance is based upon the past experiences of the observer. Over time an observer may learn where their target is most likely to appear (Geng & Behrmann, 2005), when it will appear (Olson & Chun, 2001), or what it will look like (Sha et al., 2017), without any explicit instruction. This incidentally acquired knowledge may then lead to a biasing of attention depending on the expectations the observer has developed.

While the different types of attentional influences have been studied extensively, much of the investigations and findings have concentrated only on how these manipulations impact basic measures of reaction time or accuracy derived from a dichotomous response modality. But attention has much broader effects than simply speeding up responses or assisting in choosing which of two options identify the target. Since the latter part of the 20th century, the role of attention has also been considered critical in how features, locations, and objects become bound together in our mental representations (O'Craven et al., 1999; A. M. Treisman & Gelade, 1980). Following the advent of the delayed-estimation task (Wilken & Ma, 2004), recent work has sought to expand investigations of attention beyond search tasks

and examine what types of errors are made when attempting to encode and recall a target's feature on a continuous scale, depending on how attention is directed.

For example, Golomb, L'Heureux, and Kanwisher (2014) tested feature reports following different types of top-down attentional manipulations. Of particular note, they contrasted two experiments where participants were asked to either split their attention between two possible target locations (as one was guaranteed to be the eventual target) or covertly attend to one location and then occasionally shift their covert attention to another location (the most recently cued location was always their target) before an array of colored squares appeared. Participants then reported the target color by clicking on a color wheel. The types of errors they made differed greatly depending on which of these two top-down attentional cues were used to guide attention. When attention was split between two locations, participants sometimes made feature-mixing errors, where the reported target color tended to be slightly but systematically distorted in color space, attracted towards the color that had appeared at the other pre-cued location (see Golomb, 2015, for instances in which feature distortions in the repulsion direction occur as well). However, when participants were instructed to shift their attention, a different type of feature errors – "swap errors" – were made, in which participants sometimes mistakenly reported the color of the item that appeared at the initially attended location instead of the target's color (Golomb et al., 2014). Dowd and Golomb (2019) found that these distinct types of feature errors extend to multi-feature stimuli as well, with split attention producing feature-binding errors (e.g., illusory conjunctions), whereas shifts of attention produced correlated (bound) swap errors.

Chen, Leber, and Golomb (2019) used a similar paradigm to investigate the consequences of bottom-up attentional capture. Whereas numerous prior studies had found that salient distractor cues can cause spatial attention to be temporarily captured away from the target location (see Luck et al., 2021, for a review), here the salient distractors also induced a unique combination of feature errors, a mix of both large swap errors (reporting the color of the item at the distractor location instead of the

target location) and more subtle repulsion errors (reporting a color similar to the correct target color but biased away from the color in the distractor location).

These studies have made it clear that how we represent and remember visual features depends not only on where we are attending, but *how* attention becomes attracted to different locations. This, then, raises a host of other questions. Of primary interest for the current study is the fact that attentional influences are not restricted to deterministic top-down and salient bottom-up sources. As noted earlier, attention can also be attracted to specific locations through experience-driven guidance. Experience-driven guidance has been shown to lead to quicker response times when a target is in a more likely or expected location, even if participants are not completely aware of the knowledge they have gleaned regarding these statistical regularities (Chun, 2000; Geng & Behrmann, 2005; Hutchinson & Turk-Browne, 2012; Jiang et al., 2018). If spatial attention is drawn to a nontarget location based on experience-driven guidance, will that produce a similar pattern of feature errors as when attention is drawn to that location via bottom-up stimulus-driven capture? What about other types of attentional cues that produce spatial expectations, but not certainty, e.g., explicit probabilistic cues directing top-down attention (Posner, 1980; Riggio & Kirsner, 1997)?

The present study seeks to fill this knowledge gap by examining the effects of both experience-driven and top-down probabilistic cues on feature processing and memory. Using designs similar to the tasks described above (Chen et al., 2019; Dowd & Golomb, 2019; Golomb et al., 2014), we conducted a series of three pre-registered experiments (plus additional variations reported in the supplement). All experiments involved reporting the color of one target item out of an array of four simultaneously presented stimuli. To reduce confusion, here we will refer to probabilistically indicated locations as "cued" and the target location as "probed". In Experiment 1 we first tested the impact of a well-established experience-driven attentional influence: the spatial probability cue (Geng & Behrmann, 2005; Jiang et al., 2013). By biasing the location of the target retro-probe to one “rich” (high probability)

location over the course of the experiment, we could measure what types of systematic feature report errors would be made when the target probe instead appeared in one of the low-probability locations. We used a probabilistic mixture model (Bays et al., 2009; Zhang & Luck, 2008) to analyze the color selections made along a continuous response color wheel for each participant. In this manner, we were able to test if probabilistically guiding attention via experience would lead to feature distortion (attraction or repulsion) and/or swap errors.

In fact, the spatial probability cue led to a novel pattern of feature report errors we have labeled "feature avoidance". Whereas previous studies documented swap errors made by misreporting the color presented in a salient distractor location (Chen et al., 2019) or previously attended location (Golomb et al., 2014), our spatial probability cue produced a significant amount of “reverse” swap errors. In other words, participants were significantly more likely to make swap errors reporting the color of a *control nontarget* than they were to make swap errors reporting the color of the nontarget at the high probability location. Experiments 2 and 3 follow up on this novel feature avoidance phenomenon, asking whether it is specific to experience-driven guidance or generalizes to other types of probabilistic attentional cues (Experiment 2), and what may be the mechanism behind it (Experiment 3).

General Methods: Transparency and Openness Statement

We conducted a series of eight pre-registered experiments between 2018 and 2021 (see *Table 1*); the three primary experiments are reported here in the main text, with the remaining preliminary or supplemental experiments reported in the Supplement. All experiments were preregistered on the Open Science Framework prior to data collection and approved by The Ohio State University Behavioral and Social Sciences Institutional Review Board. This includes theoretical motivation, method of participant recruitment, target sample size, exclusion criteria, experimental stimuli, task design and

procedure, and main analysis methods. Additional analyses not pre-registered are declared below as exploratory. Data for all experiments will also be available post-publication on OSF (https://osf.io/bq8yc/; https://osf.io/3se5t/). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in each study. All participants were required to be between 18-40 years of age, understand English, be capable of using a computer, and report normal or corrected to normal visual-acuity and color vision.

Experiment 1 Method

Sample

Data from 28 participants (17 female and 11 male, ages 18-20 years old) recruited from undergraduate psychology courses were analyzed for Experiment 1. Four additional participants who completed the experiment were excluded as per our pre-registered exclusion criteria (below-threshold performance in the Valid condition). Each participant received credit towards their respective psychology course for taking part in the experiment.

Our pre-registered sample size of 28 participants was derived from the results of Chen et al. (2019). We performed a series of power analyses to determine an appropriate sample size to detect both swap and feature distortion errors with 80% power. Swap errors: Chen et al.'s (2019) first experiment found a Cohen's d effect size of .791 for the probability of misreporting the salient distractor compared to the control distractor, while the second experiment reported d = .831 for the same comparison. A priori power analyses using G*Power (Faul et al., 2007) on the average of these two effect sizes, .811, (utilizing two-tailed, paired samples t-tests, an alpha of .05, and a power of 80%) resulted in an estimation of 14 participants for this swap-error analysis. Distortion errors: Chen et al.'s (2019) analyses of the shifting effect they observed in their mean target distributions found a Cohen's d effect size of .695 for the first experiment and d = .408 for the second experiment. A priori power analyses on the average of these two effect sizes for the mean shift analyses ($\bar{d}$ = .552), resulted in an

estimation of 28 participants required for this distortion-error analysis. Therefore, we set our sample size according to the more conservative estimate to ensure we would have enough power to detect both effects, if present.

Setup

Each participant was seated and placed their head against chin and forehead rests 65cm away from the monitor. The 51cm CRT monitor (resolution: 1280x1024, refresh rate: 85Hz) was color calibrated with a Minolta CS-100 colorimeter. Stimuli were generated using MATLAB (Mathworks) and the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) on a Mac computer. Eye position was recorded using an Eyelink 1000 eye-tracker (SR Research).

Procedure

At the start of every trial, a black fixation cross would appear on a grey background (RGB [127.5, 127.5, 127.5]) at the center of the screen (*Figure 1*). Once this cross had been fixated (eye position accurately maintained within a 2° radius) for a consecutive 750ms, it would change into a black dot, which in turn had to be accurately fixated for 500ms straight. If fixation was broken during this time (>2° deviation), the cross would re-appear and again require 750ms of fixation before changing to the dot again for 500ms, and this loop would continue until fixation was properly maintained for the entire 1250ms length of time. We chose to implement this two-stage fixation period in order to maximize the number of usable trials, given we would exclude any trials from analyses in which fixation was broken following this period.

Once reliable fixation was achieved, the fixation dot remained on-screen while the stimulus array was presented for 100ms. The stimulus array was four squares (each sized 2° x 2°, centered at an eccentricity of 4°), which appeared in the same upper left, upper right, lower left, and lower right positions on every trial. The color of the squares varied on every trial. The color of the upper left square was chosen randomly from 180 possible color values that were evenly distributed along a color wheel in

CIE L*a*b color space (L = 70, a = 20, b = 38, radius = 60). The colors of the squares in the upper right and lower left were then selected to be exactly 90° and -90° away in color space, direction randomly assigned on each trial. The lower right square was always 180° away in color space from the color in the upper left. Four scrambled-color masks then covered the squares for 250ms. Following the masks, the target probe appeared as a white frame at one of the four stimulus locations for 500ms. A blank delay screen appeared for 150ms before the presentation of the response screen, which consisted of a color wheel centered on the screen (diameter = 7.75°, width = .75°) displaying all 180 possible color values. Because this was a post-stimulus probe design, participants were instructed to try and remember all the colors and their respective locations, and then report the color that had appeared in the location probed by the white frame. Feedback was provided on every trial in the form of a white line on the color wheel indicating the correct color for 750ms. The color wheel was randomly rotated and flipped clockwise or counter-clockwise on any given trial. If participants broke central fixation (>2° deviation from center) during the presentation of the squares or masks, the trial was flagged for exclusion from analyses.

Our spatial probability manipulation was executed by having the target probe appear in one high probability (HP) location most often (counterbalanced across participants). This HP location is where the target would appear on 62.5% of trials over the course of the experiment; the remaining locations were each 12.5% likely to contain the target. Trials were grouped into blocks of 16, during which the target would appear in the HP location on 10 trials, with the target appearing in each of the other three locations twice per block, in a randomized order within the block. The experiment consisted of 35 blocks, which was preceded by 10 practice trials (excluded from analyses). This summed up to a total of 560 trials completed within the 1-hour time window in which we conducted this experiment. Following the completion of the 35 blocks, two exit questions were presented to evaluate the level of awareness each participant had concerning the spatial probability manipulation. The first question (EQ1) asked whether the participant perceived the target appearing most often in one consistent location, to which

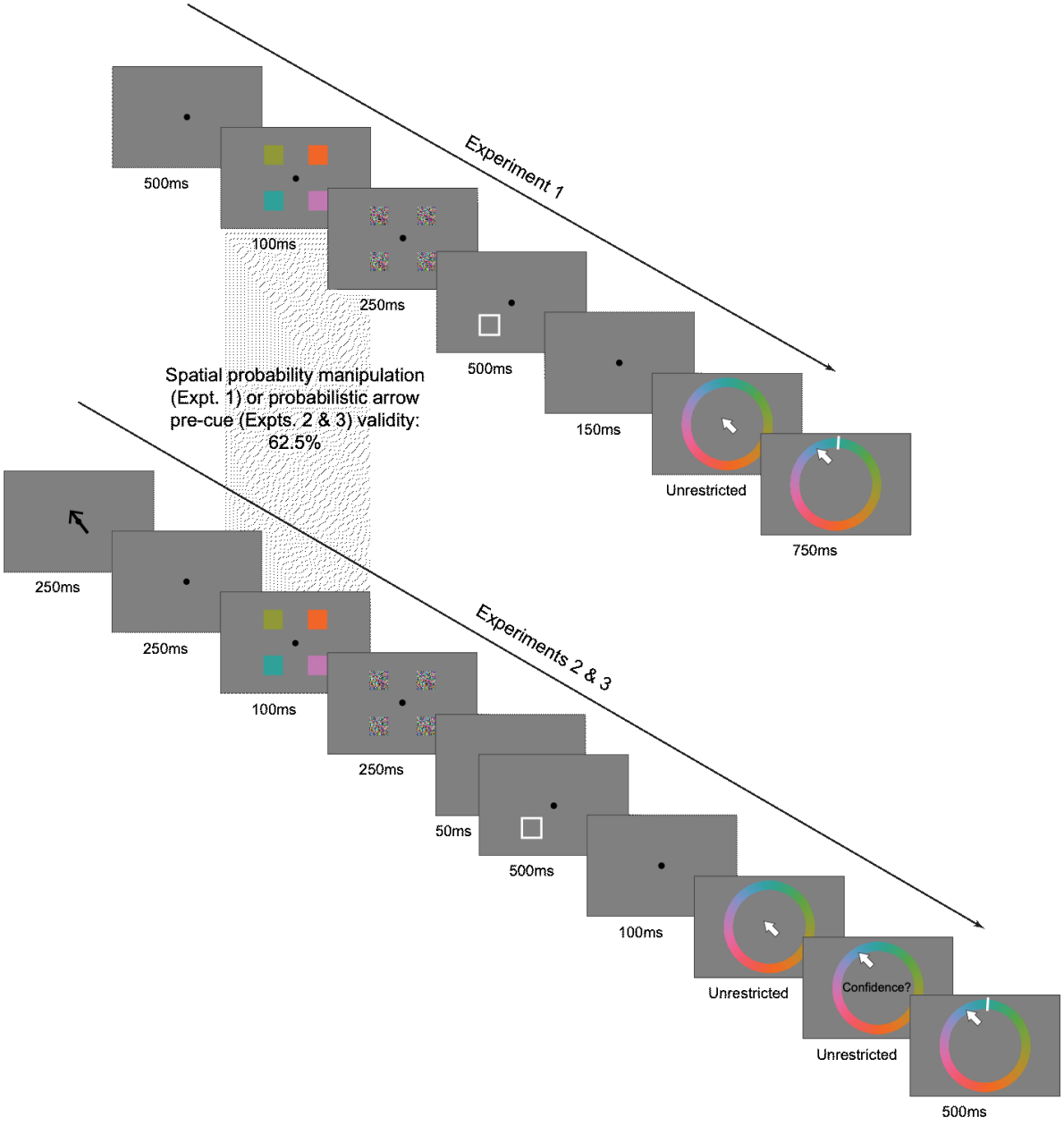

Figure 1. Experimental procedures. On every trial, participants were shown four colored squares and instructed to report the color that appeared in the location indicated by the target probe (white frame) on the subsequent color wheel. For Experiment 1, the target was more likely to appear in one high probability (HP) location across trials. For Experiments 2 and 3, an arrow pre-cue indicated the likely (HP) target location on each trial. Participants were also asked to rate the confidence in their responses, on a scale from one to four, prior to receiving feedback in Experiments 2 and 3. The color-spacing in the stimulus array for Experiment 3 was modified to ensure the squares adjacent to the target location no longer contained colors exactly opposite each other on the color wheel.

they answered “Yes” on a keyboard by pressing the ‘Y’ key, or “No” by pressing the ‘N’ key. The second question (EQ2) asked, regardless of how they responded to the prior question, to select one of the four locations in which they believe the target had been most likely to appear. Four white frames were presented in the same locations each of the colored squares had appeared in, labeled '1' through '4', and participants then responded by pressing the corresponding number key.

Analyses

We divided trials according to three conditions: target in the HP location (Valid; 350 trials per subject), target adjacent to the HP location (Critical; 140 trials per subject), or target diagonal to the HP location (data not analyzed; 70 trials per subject). The diagonal condition was included in the experiment to equate participants’ expectations across the non-HP locations, but this condition was not analyzed because the stimulus spacing and lack of corresponding control nontarget render the model fits less interpretable.

For every trial, the angular difference along the color wheel between the reported color and the target color was calculated as the response error. This error was then aligned so that the target color was at 0° and the reported color could be a maximum of ±180° away. On Critical trials, because the HP location’s color could have been located +90° or -90° from the target on the color wheel, we re-aligned the direction of response errors on the -90 trials so that the HP nontarget would always be represented at +90° and the control nontarget (the other square located adjacent to the target, diagonal to the HP location) was at -90° in our analyses. This allowed us to label response errors with a positive sign as being ‘towards’ the HP location’s color and response errors with a negative sign as ‘away’ from the HP location’s color within the Critical condition (*Figure 2A*). On half of the Valid trials (randomly selected), the sign of the response error was flipped to match the Critical trials’ realignment process and eliminate any selection confounds driven by color direction on the color wheel.

Each participant's distribution of response errors was then fit with a probabilistic mixture model (*Formula 1* for the Valid condition and *Formula 2* for the Critical condition) estimating five parameters: γ accounted for the proportion of random guesses (a uniform distribution); β_{HP} estimated the probability of misreporting the nontarget in the HP location in the Critical condition (or β_{CtlA} for one of the control nontargets in the Valid condition; i.e., a von Mises distribution centered at +90°); β_{Ctl} estimated the probability of misreporting the control nontarget in the Critical condition (β_{CtlB} in the Valid condition; i.e., a von Mises centered at -90°); and the probability of reporting the target (a von Mises distribution with a flexible mean μ, and concentration κ) was estimated by $1 - \beta_{HP} - \beta_{Ctl} - \gamma$ in the Critical condition ($1 - \beta_{CtlA} - \beta_{CtlB} - \gamma$ in the Valid condition).

Valid condition: $p(\theta) = (1 - \beta_{CtlA} - \beta_{CtlB} - \gamma)\phi_{\mu,\kappa} + \beta_{CtlA}\phi_{90°,\kappa} + \beta_{CtlB}\phi_{-90°,\kappa} + \gamma(\frac{1}{2\pi})$ (*Formula 1*)

Critical condition: $p(\theta) = (1 - \beta_{HP} - \beta_{Ctl} - \gamma)\phi_{\mu,\kappa} + \beta_{HP}\phi_{90°,\kappa} + \beta_{Ctl}\phi_{-90°,\kappa} + \gamma(\frac{1}{2\pi})$ (*Formula 2*)

The model was fit to individual participant data for each condition of interest (Valid and Critical conditions) by applying the Markov chain Monte Carlo method using MemToolbox (Suchow et al., 2013). Kolmogorov–Smirnov tests were then run on all main model fittings to ensure good fits to the raw data (all p values > .3). The best-fitting parameter estimates obtained for each subject and condition were compared in JASP software (Version 0.11.1) and MATLAB (Mathworks) using one- and two-way repeated measures ANOVAs, along with paired- and one-sample two-tailed t-tests, with significance set at $\alpha = .05$ for all tests. Our main comparisons of note involved (1) comparisons of generic performance indicators: the parameter estimates for random guessing (γ) and standard deviation (SD = $\sqrt{1/\kappa}$), and (2) comparisons of systematic feature errors: distortion errors indicated by mean shifts (μ) deviating from 0, and selective swap errors indicated by probability of nontarget reports (β_{HP} vs β_{Ctl}).

Exploratory analyses were also conducted analyzing parameter estimates by block to investigate changes in our measures of interest over time, given that that our manipulation concerns a statistical regularity that may be learned over time. Due to the mixture modeling processes requiring a large number of trials in order to fit the model, these timecourse analyses could not be conducted at the subject-level. Instead, data were aggregated across subjects for each condition, and this aggregated dataset was modeled to obtain one maximum likelihood parameter estimate for each condition per block. (For consistency across experiments, to match the number of trials going into each data point, we consolidated the 35 actual blocks in Experiment 1 into 9 blocks for this analysis, with each block containing up to 64 trials per subject.) We then conducted correlation tests (Pearson's r) and calculated the line of best fit to observe whether any significant trends over time emerged (statistics for all experiments in Table 2).

Experiment 1 Results

On average, less than 6% of trials were discarded due to fixation being broken during stimulus presentation across the 28 participants included.

Generic Performance Indicators

We first tested our basic premise that performance should be better when the target appears in the HP location (Valid condition) compared to when the target appears in a non-HP, adjacent location (Critical condition). We operationalized better performance as participants exhibiting higher precision (lower SD parameter) and/or lower likelihood of random guessing (lower γ parameter). Indeed, the guess estimates in the Valid condition (M = .08, SE = .015) were significantly lower compared to the Critical condition (M = .14, SE = .024), t(27) = -3.293, p = .003, d = -.622 (*Figure 2B*). The SD comparison also revealed a significant difference in the predicted direction between the Valid (M = 17.11, SE = .512) and Critical (M = 20.01, SE = .777) conditions, t(27) = -4.481, p < .001, d = -.847. These results support our premise; when the target appeared in the HP location, response patterns were consistent with

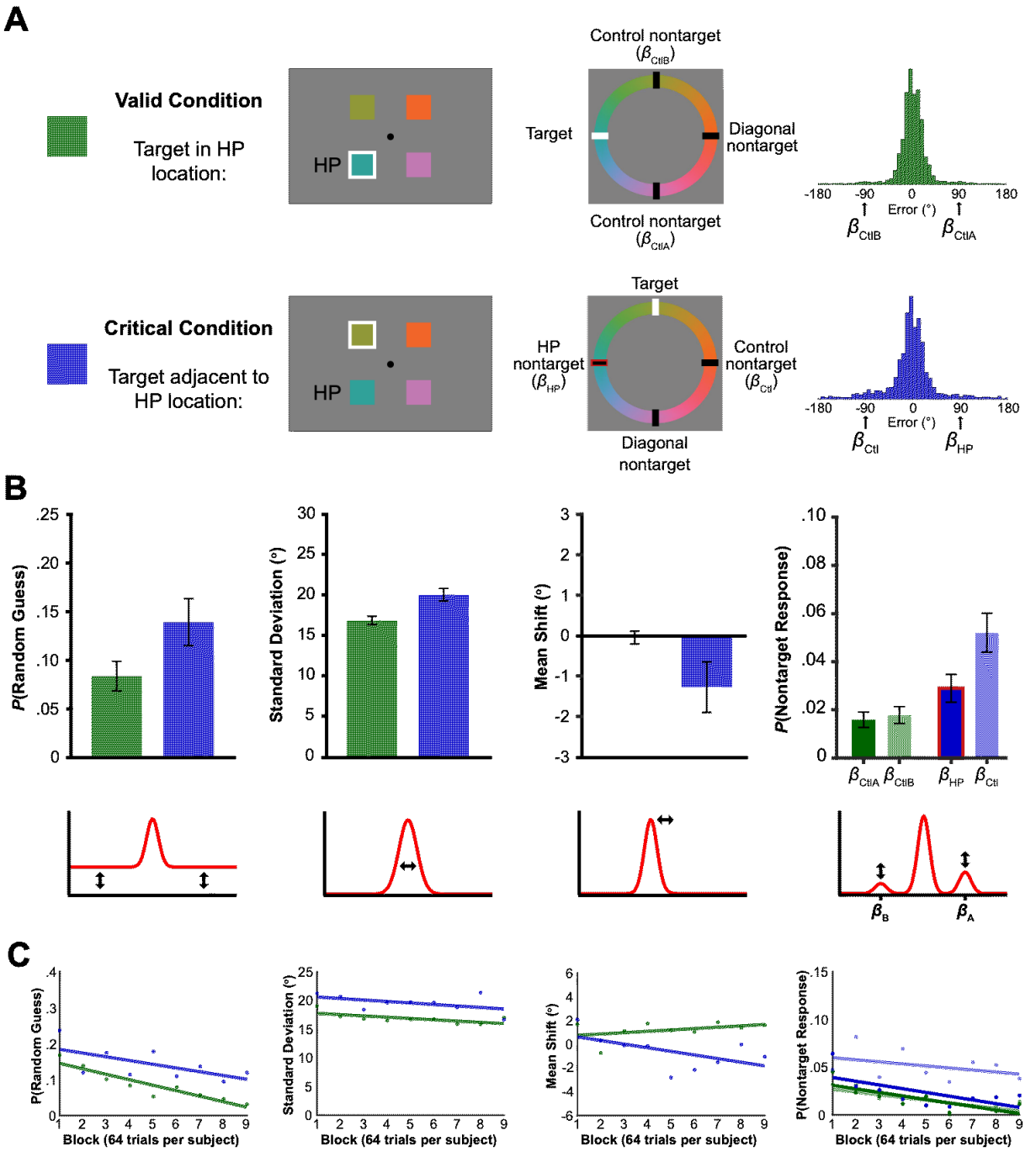

Figure 2. Experiment 1 probabilistic mixture model results. A.) Schematics of Valid and Critical conditions, illustrating example stimulus arrays based on whether the target (white frame, actually a post-cue) was in the HP location (Valid) or adjacent to it (Critical). Nontargets are shown in physical space (left) and color wheel space (right), for these illustrative examples. Response histograms collapsed across participants are shown for each condition at right, aligned as errors relative to the target (0°), and in the Critical condition, HP nontarget (+90°). B.) Mean maximum likelihood parameter estimates for: probability of random guesses (γ), SD ($\sqrt{1/\kappa}$), mean shift (mu), and probability of nontarget responses (β). Cartoons illustrating each parameter in the model are shown in

red below each plot. In the Critical condition, the nontarget in the HP location (outlined in red) is represented by β_{HP}, while β_{Ctl} represents the control nontarget; a negative mean shift indicates a biasing of target responses away from the color of the nontarget in the HP location. Error bars indicate standard error from the mean, N=28. C.) Exploratory timecourse analysis showing scatter plot of parameter estimates by experimental block (aggregated across subjects), with best-fit line. Color codes match panels A-B.

better performance, showing both less random guessing and lower SD than when the target appeared in a less likely location.

Our exploratory block analyses (*Figure 2C*) revealed a significant negative correlation between the guess parameter and block in the Valid condition, with the difference between Valid and Critical appearing to grow over time. Negative, but non-significant correlations were also measured between the SD parameter and block in both the Valid and Critical conditions (*Table 2*). These results show that participants were generally improving on these generic performance indicators over time, particularly in the Valid condition, presumably as a result of learning the spatial probability cue.

Systematic Feature Errors

Given that participants were indeed biasing their attention to the HP location, we next examined the main question: Does probabilistically guiding attention via experience lead to systematic feature errors such as feature distortion and/or swap errors when the target unexpectedly appears in an adjacent non-HP location? We first tested for feature distortion errors, testing if the mean of the target distribution in the Critical condition (M = -1.27, SE = .624) had shifted away from 0°. When collapsing across all blocks, this difference was not significant, $t(27) = -2.030$, $p = .052$, $d = -.384$, and while exploratory timecourse analysis hinted that repulsion may be growing over the course of the experiment, the negative correlation was non-significant (*Table 2*).

Next, we assessed swap errors, with a repeated measures ANOVA comparing condition (Critical vs Valid) by nontarget ($\beta_{HP\ (or\ CtlA)}$ vs $\beta_{Ctl\ (or\ CtlB)}$). We observed a significant main effect of condition, $F(1, 27) = 25.920$, $p < .001$, $\eta^2 = .227$, with more nontarget reports in the Critical condition, as well as a

significant main effect of nontarget, $F(1, 27) = 8.183$, $p = .008$, $\eta^2 = .064$. Critically, we also observed a significant condition × nontarget interaction, $F(1, 27) = 5.651$, $p = .025$, $\eta^2 = .045$. A planned t-test for the Critical condition revealed there was indeed a significant difference in the probability of reporting of the HP nontarget (M = .03, SE = .006) versus control nontarget (M = .05, SE = .008), but in the *opposite* direction anticipated, $t(27) = -2.717$, $p = .011$, $d = -.514$. This surprising result suggests participants were more likely to misreport the *control* nontarget's color than the HP nontarget's color on Critical trials, which was a response pattern not previously observed in studies that manipulated attention via top-down or bottom-up cues (Chen et al., 2019; Golomb, 2015; Golomb et al., 2014).

Interestingly, this pattern did not seem to be due to a suppression of the HP nontarget color per se. β_{HP} responses in the Critical condition were significantly more frequent than the baseline rate of nontarget responses in the Valid condition: β_{HP} vs β_{CtlA}, $t(27) = 2.337$, $p = .027$, $d = .442$ (post-hoc exploratory analysis). Rather, the observed pattern seemed due to the control nontarget attracting *more* responses than would otherwise be expected. The exploratory timecourse analysis suggests that nontarget responses generally tended to decrease over time, significantly so in the Valid condition (*Table 2*).

Exit Questions

In post-experiment exit questions, participants displayed a relatively high level of awareness for the spatial probability cue. 17 of the 28 participants answered "Yes" to noticing a target-location bias in EQ1. 24/28 (86%) correctly identified their HP location (EQ2), which was significantly higher than chance (25%), according to a binomial test ($p < .001$). We sorted participants into those with 'explicit awareness' (defined as participants who answered "Yes" to EQ1 and answered EQ2 correctly; N = 16) and those without, and found that this factor had no significant impact on the main results (no significant interactions between explicit awareness and guessing, precision, or swap rate: all p-values > .42, exploratory analysis).

Experiment 1 Discussion

The generic performance indicators (guess rate and SD) confirmed that our spatial probability cue was effective at guiding attention in this context. Participants were less likely to make random guesses and were more precise in reporting the correct color when the target appeared in the HP location (Valid condition) versus a location adjacent to the HP location (Critical condition). Combined with the timecourse (block-wise) analysis, this suggests that our spatial probability cue guided participants' covert attention to the location where they learned over time to expect the target to appear, based on their experience.

Participants also made systematic feature errors on Critical trials. However, instead of observing swapping errors where participants selectively misreported the feature in the anticipated/attended location (*HP* nontarget), which would have resembled the effects of top-down or bottom-up influences (Chen et al., 2019; Golomb et al., 2014), we observed a different type of error: a higher likelihood to swap the color of the *control* nontarget. This suggests a unique effect of our experience-dependent cue, as if participants *avoided* reporting the feature in the anticipated HP target location on trials when the target appeared elsewhere. We propose that spatial attention is attracted in advance to the HP location, so effectively that on some trials *only* the item in that HP location is reliably encoded (and/or a strong color-location binding is made only for that one HP item). On trials when the target probe unexpectedly appears elsewhere (i.e., Critical trials), participants may not be able to correctly reconstruct the actual target color – but perhaps they have some awareness that the strongly encoded color from the HP location was *not* the probed target. Thus, the observed pattern of response errors may represent a response strategy to avoid that one color they know is not the target color. We label this tendency to avoid an attended feature value "feature avoidance".

Notably, this avoidance is not the same as feature suppression. In other words, the results are not consistent with the representation of the HP item’s color being suppressed during perception or

memory (the rate of selecting the HP nontarget was still above the baseline rate of selecting nontargets in the Valid condition). Rather, the effect seems more consistent with a type of strategic guessing and/or location binding error that manifests when expectations based on probabilistic spatial attention are violated.

In Experiments 2 and 3, we further investigate this feature avoidance effect. In Experiment 2, we first assess how reliable this feature avoidance effect is across contexts, specifically whether it is a result of implicit experience-driven spatial probability learning or might be a product of probabilistic attentional cues more generally. We also incorporate confidence ratings after each response in order to test our hypothesis that feature avoidance is the result of a response strategy stemming from uncertainty. We further probe the nature of the feature avoidance effect in Experiment 3 (and supplemental experiment S5) by varying the spacing of items in color space to differentiate different potential types of feature avoidance.

Experiment 2

Is the novel feature avoidance effect reported in Experiment 1 a result of manipulating experience-driven attention specifically? Or is it the probabilistic nature of the cue that is the driving factor in producing this pattern of feature errors? To address this question, we conducted a second experiment in which we replaced the experience-driven spatial probability cue with a top-down spatial probability cue. Here we used a prototypical top-down attentional influence: a central arrow pre-cue (Posner, 1980; Riggio & Kirsner, 1997). Note that in contrast to the top-down attentional manipulations studied previously using similar continuous-report tasks (Dowd & Golomb, 2019; Golomb, 2015; Golomb et al., 2014), the top-down cue in Experiment 2 was *probabilistic*. In other words, in the prior studies, a spatial pre-cue indicated the target location with 100% certainty, whereas in Experiment 2, the arrow pre-cue indicated the probable target (HP) location. We matched this probability (62.5% Valid trials) to

that of the spatial probability cue from Experiment 1. If feature avoidance is a specific consequence of experience-driven attention, then in Experiment 2, we would expect to see more of a standard pattern of swap errors to the HP location, similar to when covert attention is intentionally shifted from one location to another (Golomb et al., 2014). On the other hand, if feature avoidance stems from probabilistic attentional guidance, we would expect to observe a similar signature of feature avoidance as in Experiment 1: an increased tendency to select the color of the control nontarget on Critical trials. The inclusion of a confidence measurement after each response also allows us to discern whether feature avoidance errors, if observed, are made with relatively high confidence, suggesting perceptual binding errors or shifts in underlying memory representations, versus low confidence, suggesting a response based on uncertainty.

Experiment 2 Method

Participants

A new set of 28 naïve participants (20 female and 8 male, ages 18-29 years old) were recruited, and either received course credit or $10 for their time. Two additional participants who completed the experiment were excluded, one due to a programming error and the other for failing to follow instructions and only reporting where the arrow cued instead of the target probe.

Setup

Each participant was seated and placed their head against chin and forehead rests 60cm away from the monitor. The 62cm LCD monitor's resolution was adjusted to display a 4x3 presentation window (resolution: 1280x960, refresh rate: 200Hz) and was color calibrated with a Minolta CS-100 colorimeter. Stimuli were generated using MATLAB (Mathworks) and the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) on a Windows computer. Eye position was recorded using an Eyelink 1000 eye-tracker (SR Research).

Procedure

The procedure closely followed the design of Experiment 1, with the following changes. The spatial probability manipulation was removed, so that each of the four possible locations was now equally likely to contain the target across the experiment. Our attentional manipulation was instead based on a central, black arrow cue (length = 1.56°, stroke = .22°) that appeared 500ms prior to the stimulus array of colored squares. The arrow appeared at fixation for 250ms and pointed towards one of the four possible stimulus locations, followed by a 250ms fixation delay. This 500ms total time between the onset of the arrow cue and the onset of the stimulus array was chosen to ensure sufficient time for the endogenous cue to be cognitively processed (Müller & Rabbitt, 1989). Participants were told that the arrow would indicate the same location as the target post-probe on most of the trials, but not all of the trials. It was stressed that the task was to encode the stimuli and report the color of the item that appeared wherever the target probe subsequently indicated, regardless of where the arrow had pointed. Due to the additional time added by presentation of the arrow pre-cue, minimum fixation time in-between trials was reduced from 1250ms to 1000ms and the feedback time was reduced from 750ms to 500ms to try and maintain a similar experiment duration.

Eight blocks of 64 trials were conducted. The arrow matched the target location on 62.5% of trials (Valid condition) and indicated one of the other 3 locations 12.5% of the time each. As in Experiment 1, the Critical condition was defined as the target post-probe indicating a location adjacent to the HP location (pre-cue arrow), and β_{HP} in the mixture model was assigned to the HP nontarget item (aligned to +90°).

To get a better sense of participants' confidence in their responses and test whether feature avoidance responses, if present, are made with relatively low or high confidence, a Likert-style confidence rating was also obtained after every response, prior to presenting the feedback display. The text, "Confidence?" appeared after a selection was made on the color wheel, and participants were instructed to press the "1" key when completely guessing, the "4" key when almost certain in their

response, and "2" or "3" when they felt somewhere in-between. One exit question was included to query the perceived reliability of the arrow on a scale from 10% to 90% reliable. Other than the described changes, the procedure was identical to Experiment 1.

Experiment 2 Results

On average, less than 5% of trials were discarded due to fixation being broken during stimulus presentation across the 28 participants included in the following analyses.

Generic Performance Indicators

We first tested our basic premise that performance should be better when the target appears in the HP location (Valid condition) compared to an adjacent location (Critical condition). The same operational definition for better performance was used as before; lower likelihood of random guessing and/or lower SD. As with the spatial probability cue, the arrow pre-cue appeared effective at guiding attention to the HP location. Random guess rates were significantly lower on Valid trials (M = .07, SE = .012) relative to Critical trials (M = .20, SE = .026), $t(27) = -6.091$, $p < .001$, $d = -1.151$ (*Figure 3B*). The performance advantage for Valid trials was further supported by the significantly smaller SD estimate (M = 15.71, SE = .590) compared to Critical trials (M = 21.77, SE = 1.061), $t(27) = -7.757$, $p < .001$, $d = -1.466$. Overall, the lower likelihood of guessing and increased response precision in the Valid condition showed that better performance was elicited by a valid than invalid arrow cue. Therefore, we could proceed with confidence that our premise was met, and participants were indeed biasing their attention towards the location indicated by the arrow.

Our exploratory block analyses (*Figure 3C*) revealed significant negative correlations in the Valid condition between guess rate and block and between SD and block. In the Critical condition, neither guess rate nor SD were found to correlate significantly with block. These results suggest that participants' attention was being increasingly guided to the HP location by the arrow as the experiment progressed.

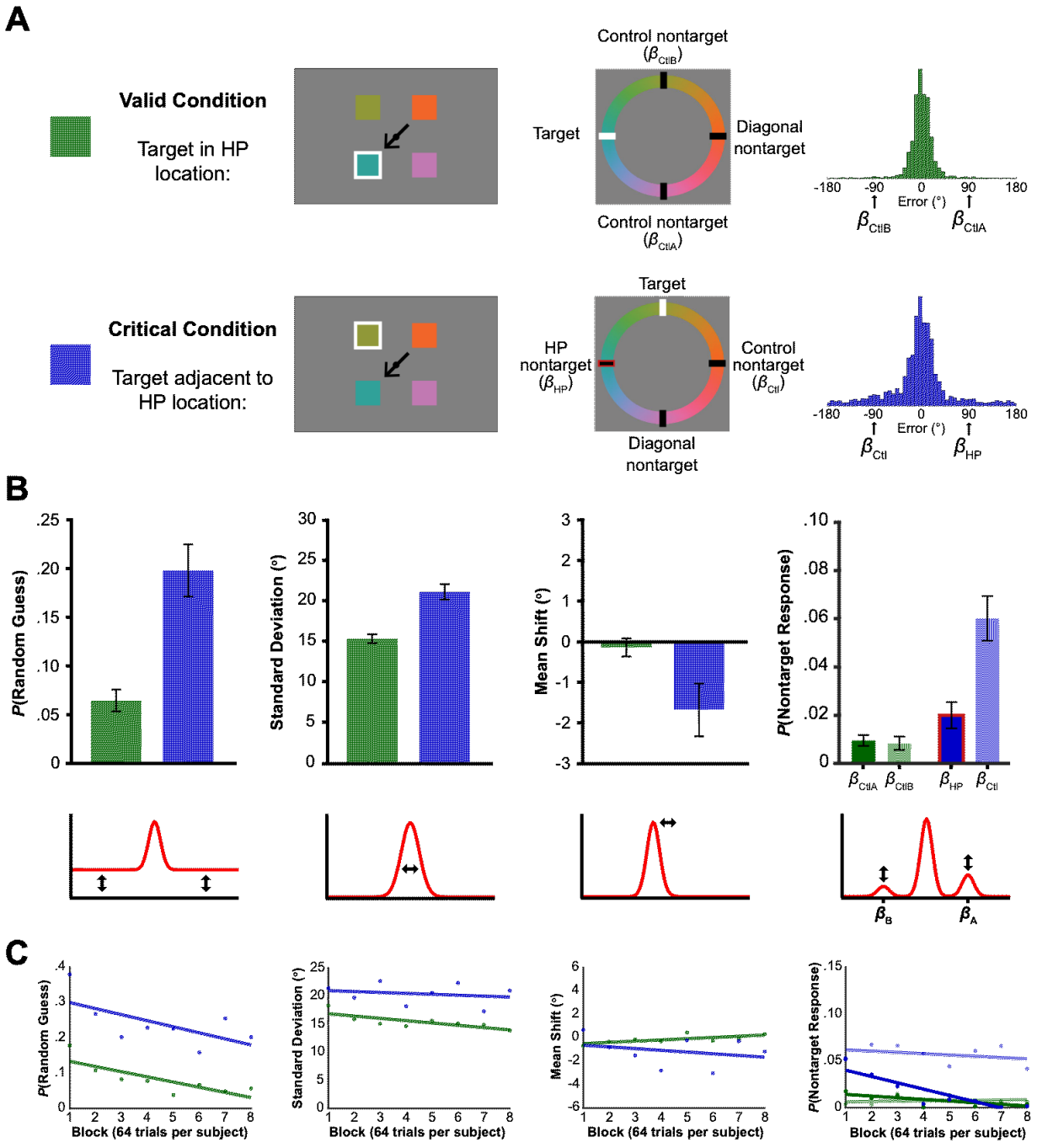

Figure 3. Experiment 2 probabilistic mixture model results. A.) Schematics of Valid and Critical conditions, illustrating example stimulus arrays based on whether the target (white frame, actually a post-cue) was in the HP (arrow pre-cued) location (Valid) or adjacent to it (Critical). Nontargets are labeled in physical space (left) and color wheel space (right), for these illustrative examples. Response histograms collapsed across participants are shown for each condition at right, aligned as errors relative to the target (0°), and in the Critical condition, HP nontarget (+90°). B.) Mean maximum likelihood parameter estimates for: probability of random guesses (γ), SD ($\sqrt{1/\kappa}$),

mean shift (mu), and probability of nontarget responses (β). Cartoons illustrating each parameter in the model are shown in red below each plot. In the Critical condition, the nontarget in the HP location (outlined in red) is represented by β_{HP}, while β_{Ctl} represents the control nontarget; a negative mean shift indicates a biasing of target responses away from the color of the nontarget in the HP location. Error bars indicate standard error from the mean, N=27. C.) Exploratory timecourse analysis showing scatter plot of parameter estimates by experimental block (aggregated across subjects), with best-fit line. Color codes match panels A-B.

Systematic Feature Errors

Next, we examined feature distortion (mean shift) and swap (nontarget responses) errors to test whether the probabilistic arrow pre-cue would elicit response errors more similar to the spatial probability cue from Experiment 1, or deterministic top-down cues from previous work (Golomb et al., 2014). The results revealed a similar feature avoidance effect as Experiment 1.

The arrow pre-cue elicited a significant mean shift away from the arrow-cued nontarget color (M = -2.14, SE = .783), t(27) = -2.733, p = .011, d = -.517 . Our exploratory block analysis showed this repulsion effect as appearing relatively constant over time, as no significant correlation was measured between mean shift and block in the Critical condition (*Figure 3C*).

We then analyzed swap errors using a repeated measures ANOVA comparing condition (Critical vs Valid) and nontarget ($\beta_{HP\ (or\ CtlA)}$ vs $\beta_{Ctl\ (or\ CtlB)}$). A main effect of condition was found to be significant, $F(1, 27) = 29.813$, $p < .001$, $\eta^2 = .214$, but a main effect of nontarget was not, $F(1, 27) = 3.588$, $p = .069$, $\eta^2 = .035$. Critically, a significant interaction (condition × nontarget) was observed, $F(1, 27) = 4.244$, $p = .049$, $\eta^2 = .040$, with greater swaps to the control nontarget than the HP nontarget in the Critical condition. The exploratory block analysis revealed that within the Critical condition, the likelihood of misreporting the HP nontarget (β_{HP}) significantly decreased over time, while no significant correlation was observed between the likelihood of misreporting the control nontarget (β_{Ctl}) and block (*Table 2*).

Surprisingly, despite the significant interaction and large numerical difference between the β_{HP} (M = .03, SE = .011) and β_{Ctl} (M = .06, SE = .010) parameters, the difference between β_{HP} and β_{Ctl} was not

statistically significant, t(27) = -1.992, p = .057, d = -.377. Upon closer examination, we noticed that one participant seemed to exhibit an extreme, outlier pattern that diverged from the norm of the group on several measures. In particular, this participant reported the HP nontarget color (β_{HP}) on nearly one-third of trials. While this value technically did not exceed our pre-registered exclusion criteria, it exceeded the group mean by >4 SD. Moreover, this same participant was also a statistical outlier (>2.5 SD from the sample mean) on other measures, including the Critical condition SD and mean shift model estimates, and their exit question response was also at the extreme end. Since we had not pre-registered any outlier criteria, we analyzed our data both with and without this outlier participant. When excluding this participant, the guess rate (Valid vs Critical, t(26) = -5.848, p < .001, d = -1.125), SD estimate (Valid vs Critical, t(26) = -7.632, p < .001, d = -1.469), and mean shift (Critical vs 0, t(26) = -2.567, p = .016, d = -.494) comparisons had no change in significance or direction as a result of the outlier's exclusion. The ANOVA results for the swap errors were also consistent, and the t-test directly comparing the β_{HP} (M = .02, SE = .006) and β_{Ctl} (M = .07, SE = .010) misreports reflected a significant difference, t(26) = -4.368, p < .001, d = -.841 – i.e., strong feature avoidance – across the remaining 27 participants when the outlier was excluded.

Confidence Ratings

We conducted an analysis comparing the average confidence scores (*Figure 4*) participants gave when making different types of responses. Correct target responses were defined as color errors close to 0° [±30°], HP nontarget responses as color errors close to +90° [±30°], and control nontarget responses (feature avoidance errors as color errors close to -90° [±30°]); see *Figure 4A*. Analogous bins were defined for each condition (Valid vs Critical). 23 participants made at least one response within a ±30° range of all of the stimuli of interest and could therefore be included in this set of analyses. A repeated-measures ANOVA revealed a significant main effect of condition, F(1, 22) = 5.792, p = .025, η^2 = .022, indicating participants were indeed more confident on Valid trials, which complements the

evidence from our general performance advantage measures (guess rate, SD) that the arrow was indeed effectively guiding attention. A main effect of response also emerged, $F(2, 44) = 77.916$, $p < .001$, $\eta^2 = .497$, as well as a significant interaction (condition × response), $F(2, 44) = 8.936$, $p < .001$, $\eta^2 = .073$. *Figure 4B* depicts the mean confidence ratings for each type of response in the Critical condition. Participants were most confident when correctly reporting the target color, and least confident when making feature avoidance errors (control nontarget responses). A post-hoc comparison revealed confidence ratings for feature avoidance errors to the control nontarget were significantly lower relative to ratings for HP nontarget responses in the Critical condition, $t(22) = -2.934$, $p = .008$, $d = .612$ (Holm-

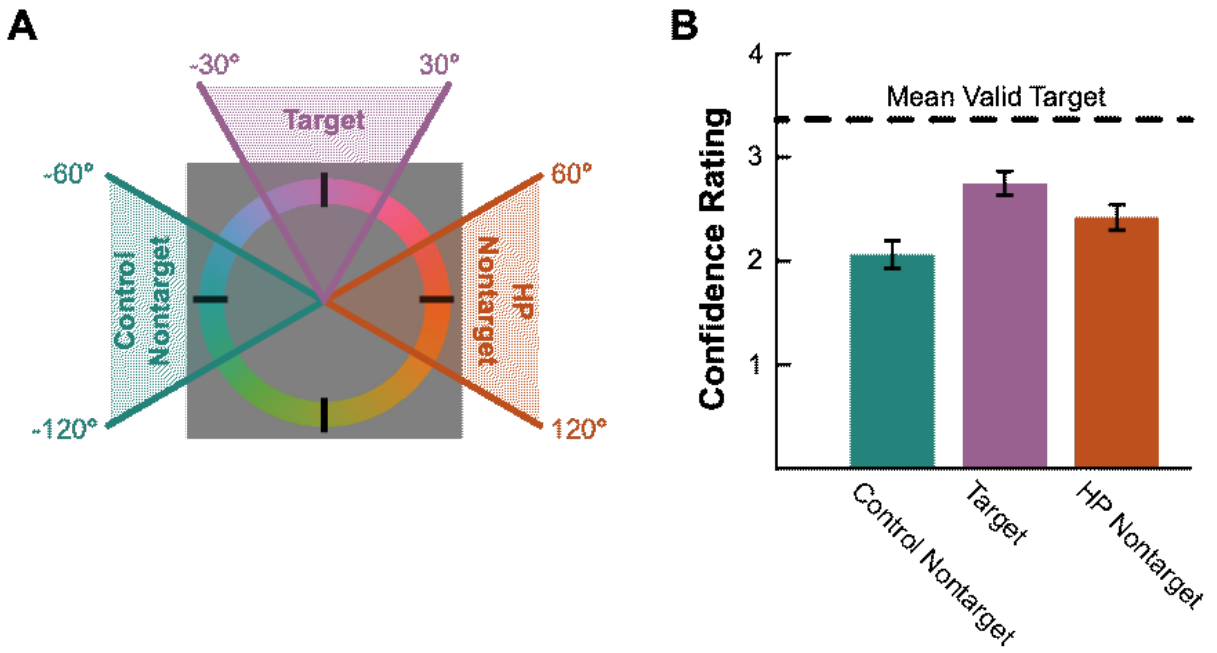

Figure 4. Experiment 2 confidence ratings by response type. A.) Schematic showing the definition of each response type in the Critical condition for this analysis. Example target and nontarget items are indicated on the color wheel with black notches, with the Target item aligned to 0°. Correct target responses (purple wedge) were defined as responses within ±30° error from 0°, HP nontarget responses (orange wedge) as responses within ±30° of the HP nontarget (+90°), and control nontarget responses (teal wedge) as responses within ±30° of the control nontarget (-90°). B.) Confidence results. Bars plot the mean confidence ratings for each type of Critical condition response, corresponding to the definitions in A. Error bars indicate standard error from the mean. Black dashed line shows the average confidence rating for correct target responses in the Valid condition as a reference.

Bonferroni corrected). This suggests that, even though participants were more likely to select the control nontarget, they were less confident when choosing the control nontarget relative to when they selected the target or HP nontarget. These results further support the idea that feature avoidance represents a response strategy when participants are uncertain of the target's color, in contrast to the high-confidence swap errors seen in Chen et al. (2019) or a perceptual phenomenon such as illusory conjunctions.

Exit Questions

Following the completion of the experiment, we asked participants to rate how often the arrow seemed to match the location of the target probe. The mean response across all participants was 59.3%, which did not significantly differ from the true value of 62.5%, t(26) = 1.297, p = .206, d = .249. We also conducted correlational analyses to test if the feature avoidance effect (Critical condition $\beta_{HP} - \beta_{Ctl}$) could be related to how reliable participants perceived the arrow to be. No significant correlation was found, r(25) = .014, p = .944, suggesting that how trustworthy participants perceived the arrow to be did not modulate the magnitude of feature avoidance.

Experiment 2 Discussion

These first two experiments have shown that feature avoidance can be elicited via both a spatial probability cue and a probabilistic arrow pre-cue. Despite the surface-level differences between these two cues of spatial attention, the results indicate they both lead to a response pattern in which the color of a control nontarget is more likely to be mistakenly selected than a nontarget they had biased their attention towards. Indeed, a repeated measures ANOVA with experiment (1 vs 2) as a between subjects factor when comparing condition and nontarget revealed no significant main effect or interactions with experiment (all p-values > .07).

The confidence data from Experiment 2 also revealed that participants were less confident when making feature avoidance errors than other types of responses. This further supports the notion

that feature avoidance stems from a response strategy evoked when probabilistic cues guide attention to what turns out to be a nontarget location. When the target probe did not match the HP location participants were selectively attending based on the pre-cue, participants seemed aware they had a poorer mental representation of the target color. We suggest that this violation of spatial expectations and associated uncertainty leads to a response strategy where participants relay on incomplete or poorer quality memory representations. When confidence about the true target color is particularly low, participants adopt the feature avoidance strategy, resulting in increased likelihood to select the color opposite the HP nontarget color, i.e., the control nontarget. Experiment 3 further probes the nature of feature avoidance and the content of memory representations under probabilistic spatial attention, testing two possible sources of information that could have been drawing responses toward the control nontargets in Experiment 1 and 2.

Experiment 3 Introduction

Is feature avoidance an increase in the probability of reporting the control nontarget specifically, or is it more of a general avoidance of reporting the invalidly cued HP nontarget's color on that trial? In other words, are participants avoiding the color that appeared in the HP location by picking a *color* maximally different during the color wheel report? Or is there actually something about the control nontarget item that has produced a specific retrievable representation and/or confusability with the target in working memory? In Experiments 1 and 2, these two possible types of feature avoidance could not be differentiated. Because the four colors in the stimulus array were evenly spaced around the color wheel, the control nontarget color was always directly opposite (180°) in color space from the HP nontarget. Therefore, we could not assess whether participants were selecting the control nontarget on Critical trials because they remembered seeing it in the array, or because they were clicking on the section of the color wheel most different from the HP nontarget color.

To resolve this confound, in Experiment 3 we modified the color spacing of the stimulus array. By having the control nontarget's color no longer appearing directly opposite the HP nontarget's color on the color wheel, we could compare whether participants were more likely to select the color of the control nontarget or the color maximally different from the HP nontarget's color. We opted for the probabilistic arrow pre-cue for Experiment 3 as it appeared to be a more efficient attentional cue (see *supplemental experiment S3)*, but we did also conduct an analogous experiment using the spatial probability cue (see *supplemental experiment S5*).

Experiment 3 Method

Participants

Due to a global pandemic, Experiment 3 was converted to online delivery and participants completed the experiment on their own computers. A sample of 56 naïve participants (19 female, 35 male, 2 non-binary, ages 18-40 years old) was recruited from either The Ohio State University or Amazon Mechanical Turk, and received course credit or $10 for their time, respectively. The pre-registered sample size was doubled from Experiments 1-2 to ensure enough power to distinguish between the two competing accounts for the feature avoidance effect. An additional 27 participants were excluded, 24 for not meeting our pre-registered exclusion criteria and 3 for technical errors possibly attributable to their personal computers; we note that higher exclusion rates are more common with online data collection.

Ten different participants were also run in an initial version of this experiment with a shorter presentation time that resulted in very low performance (see **Procedure**); these participants are not included in the analyses.

Setup

Experiment 3 was programmed in JavaScript in order to allow anyone with the associated link to run in the experiment in their own browser. Only desktop versions of the Google Chrome browser were

allowed to run the experiment, which participants were automatically notified of if they attempted to open the program in some other manner. While the colors were calibrated to be isoluminant on author WNM's computer in the same manner as the previous experiments, consistency across the various personal computers used by participants was impossible to ensure.

The colored squares appeared as solid 100px x 100px images at an eccentricity of 212px from the center of the fixation point to the center of a square. Stimuli were presented on a white background and the target probe was a black frame.

Procedure

The procedure closely followed that of Experiment 2, with a few notable changes owing to our novel manipulation and transition to online delivery. The main design change from Experiment 2 was the modification of the color-spacing in the stimulus array. The adjacent nontargets were selected to be +120° and -120° (instead of +90° and -90°) from the target on the color wheel. The nontarget diagonal to the target remained at 180° on the color wheel. Therefore, in the Critical condition the HP nontarget color was at +120° and the control nontarget color was at -120°. This meant that the color maximally different from the HP nontarget was at -60° (*Figure 5A*), allowing us to differentiate whether participants were more likely to mistakenly select the control nontarget seen in the array (errors centered on -120°) or the maximally different color on the color wheel (errors centered on -60°).

Due to the remote, online delivery of Experiment 3, eye-tracking was no longer conducted to ensure central fixation was maintained. Regardless, participants were still instructed to keep their eyes on a central fixation dot whenever it was present. Additionally, the presentation time for the stimulus array was doubled from Experiment 2 to 200ms. This change was made due to an overall high amount of poor performance observed after running ten initial participants in the online version at the intended 100ms presentation time. Other than the changes listed here, the experiment was designed to closely mirror Experiment 2.

Analyses

A third nontarget parameter was added to our probabilistic mixture model centered at -60° error, indicating the color opposite the HP nontarget (β_{HPopp} in the Critical Condition, β_{CtlAopp} in the Valid Condition). $\beta_{\text{HP (CtlA)}}$ and $\beta_{\text{Ctl (CtlB)}}$ were now located at +120° and -120°, respectively, indicating the probability of reporting the nontarget items in the display (*Formula 3* for the Valid condition and *Formula 4* for the Critical condition).

$$\text{Valid condition: } p(\theta) = (1 - \beta_{\text{CtlA}} - \beta_{\text{CtlB}} - \beta_{\text{CtlAopp}} - \gamma)\phi_{\mu,\kappa} + \beta_{\text{CtlA}}\phi_{120°,\kappa} + \beta_{\text{CtlB}}\phi_{-120°,\kappa} + \beta_{\text{CtlAopp}}\phi_{-60°,\kappa} + \gamma(\tfrac{1}{2\pi})$$

(*Formula 3*)

$$\text{Critical condition: } p(\theta) = (1 - \beta_{\text{HP}} - \beta_{\text{Ctl}} - \beta_{\text{HPopp}} - \gamma)\phi_{\mu,\kappa} + \beta_{\text{HP}}\phi_{120°,\kappa} + \beta_{\text{Ctl}}\phi_{-120°,\kappa} + \beta_{\text{HPopp}}\phi_{-60°,\kappa} + \gamma(\tfrac{1}{2\pi})$$

(*Formula 4*)

Our main comparison of interest was whether there would be a higher probability for mistakenly selecting the control nontarget color (β_{Ctl}) versus the unseen HP-opposite color (β_{HPopp}) in the Critical condition. Since both types of errors reflect feature avoidance, we first confirmed that the sum of β_{Ctl} and β_{HPopp} was greater than β_{HP}. We also conducted a Critical condition model comparison between two additional probabilistic mixture models that each only included two nontarget parameters, either $\beta_{\text{HP}}\phi_{120°,\kappa}$ and $\beta_{\text{Ctl}}\phi_{-120°,\kappa}$ or $\beta_{\text{HP}}\phi_{120°,\kappa}$ and $\beta_{\text{HPopp}}\phi_{-60°,\kappa}$. Individual participant Critical condition data were re-fit to these additional two models, which were then compared for goodness of fit according to the Bayesian information criterion using MemToolbox (Suchow et al., 2013). (Although confidence ratings were collected as in Experiment 2, the more complicated design of Experiment 3 is not particularly conducive to analyzing these data, and as we did not propose any direct hypotheses relating to confidence, we do not report those results here.)

Experiment 3 Results

Generic Performance Indicators

The arrow appeared effective at guiding spatial attention in this online format. Guess rates in the Valid condition (M = .10, SE = .015) were significantly lower relative to the Critical condition (M = .20, SE = .024), t(55) = -4.887, p < .001, d = -.653. Response SD estimates were also smaller in the Valid (M = 15.51, SE = .412), compared to Critical (M = 22.95, SE = 1.415), condition, t(55) = -5.559, p < .001, d = -.743.

Feature Avoidance Errors (Misreport parameters)

The main goal of Experiment 3 was to elucidate the driving factor(s) behind the feature avoidance effect by teasing apart different sources of avoidance errors on Critical trials (β_{Ctl} vs β_{HPopp}).

First, to confirm that feature avoidance was still present overall in this online format, we summed β_{Ctl} and β_{HPopp} because either of these types of responses would constitute feature avoidance, and we conducted a repeated measures ANOVA comparing this summed measure to $\beta_{HP\ (or\ CtlA)}$ errors for Valid vs Critical conditions. Replicating the earlier experiments, we found a significant main effect of condition, $F(1, 55) = 42.383$, $p < .001$, $\eta^2 = .180$, and nontarget, $F(1, 55) = 19.011$, $p < .001$, $\eta^2 = .080$, and a significant interaction (condition × nontarget), $F(1, 55) = 11.719$, $p = .001$, $\eta^2 = .048$. The follow-up planned paired t-test between the β_{HP} (M = .03, SE = .007) and the sum of β_{Ctl} and β_{HPopp} (M = .09, SE = .014) in the Critical condition was also significant, t(55) = -3.925, p < .001, d = -.524, which provides strong evidence that participants in this online experiment were making feature avoidance errors of one or both types.

Next, we tested whether these avoidance errors were more likely to stem from selecting the color of the control nontarget item in the display (β_{Ctl}) or selecting the maximally different HP-opposite color on the color wheel (β_{HPopp}). The average probability of reporting the control nontarget's color (β_{Ctl}: M = .06, SE = .013) was about double that of the maximally different (but unseen) color (β_{HPopp}: M = .03, SE = .005) in the Critical condition. However, the difference between these two probabilities was not

significant, t(55) = 1.929, p = .059, d = .258. We also conducted a Critical condition model comparison comparing two variations of mixture models each containing only two nontarget distributions: one which only included β_{HP} and β_{Ctl} parameters, and the other which only included β_{HP} and β_{HPopp} parameters. Overall, there was a slight preference (34/56 participants with lower BICs) for the β_{HP} β_{Ctl} model relative to the β_{HP} β_{HPopp} model. The exploratory timecourse analyses revealed significant negative correlations in the Critical condition for both β_{HP} responses and β_{Ctl} responses, but not β_{HPopp} responses. Interestingly, this pattern suggests that participants became less likely to misreport one of the actual nontarget items over time in the Critical condition, but there was no significant change over time in their tendency to select the maximally different (unseen) color.

Mean shift parameter

In contrast to Experiments 1 and 2, the mean of the target distribution in Experiment 3 (M = 1.57, SE = .650) was shifted in the opposite direction, towards the HP nontarget in the Critical condition, t(55) = 2.408, p = .019, d = .322. However, the interpretability of this attraction effect is questionable given the modifications made to our mixture model. In order to separate misreports made to the control nontarget (-120°) and the color maximally different from the HP nontarget (-60°), we added a third nontarget distribution in the model, such that β_{HP}, β_{Ctl}, and β_{HPopp} distributions were centered on +120°, -120°, and -60°, respectively. The presence of β_{HPopp} in the model, located more closely to the target at -60°, may have created an artificial asymmetry and led some negatively-shifted errors to be attributed to this nontarget distribution, while there was no symmetric distribution on the positive side. The mean shift parameter of the target distribution therefore may be problematic to interpret.

An unmodeled measure of the raw mean of the entire response distribution (without attempting to attribute errors to one source or another), confirmed that participants were significantly more likely to make feature errors in the negative (repulsion/avoidance) direction than positive direction in all three experiments (Experiment 1, t(27) = -3.004, p = .006, d = -.568, Experiment 2, t(27) =

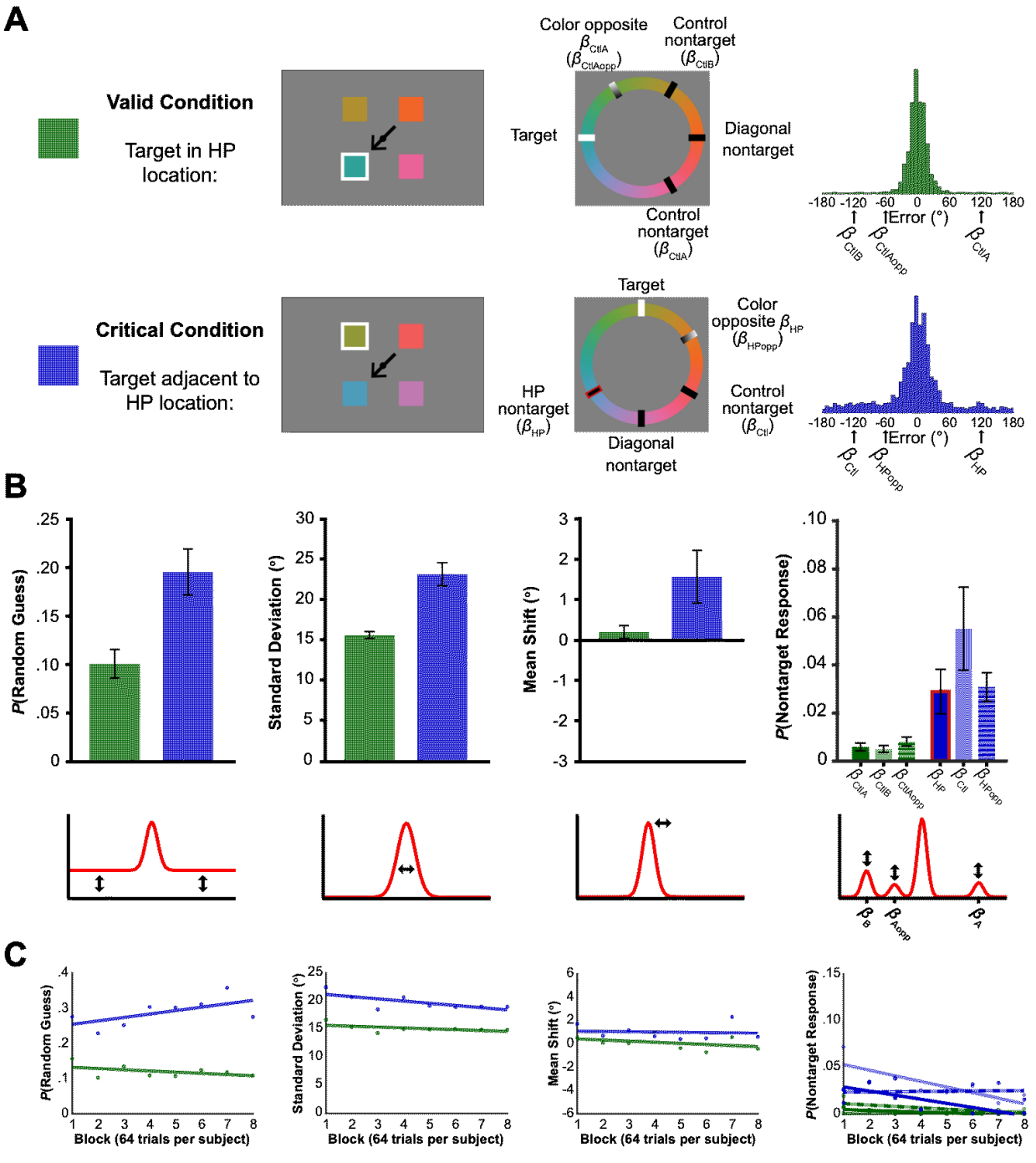

Figure 5. Experiment 3 probabilistic mixture model results. A.) Schematics of Valid and Critical conditions, illustrating example stimulus arrays based on whether the target (white frame, actually a post-cue) was in the HP (arrow pre-cued) location (Valid) or adjacent to it (Critical). Nontargets are labeled in physical space (left) and color wheel space (right), for these illustrative examples. Response histograms collapsed across participants are shown for each condition at right, aligned as errors relative to the target (0°), and in the Critical condition, HP nontarget (+90°). B.) Mean maximum likelihood parameter estimates for: probability of random guesses (γ), SD ($\sqrt{1/\kappa}$),

mean shift (mu), and probability of nontarget responses (β). Cartoons illustrating each parameter in the model are shown in red below each plot. In the Critical condition, the nontarget in the HP location (outlined in red) is represented by β_{HP}, while β_{Ctl} represents the control nontarget and β_{HPopp} represents the color opposite the HP nontarget in color space; a negative mean shift indicates a biasing of target responses away from the color of the nontarget in the HP location. Error bars indicate standard error from the mean, N=56. C.) Exploratory timecourse analysis showing scatter plot of parameter estimates by experimental block (aggregated across subjects), with best-fit line. Color codes match panels A-B.

-2.751, p = .010, d = -.520, Experiment 3, t(55) = -2.415, p = .019, d = -.323).

Exit Questions

As in Experiment 2, at the conclusion of the experiment we asked participants to rate how reliable they perceived the arrow pre-cue to have been on a scale from 10% to 90%. Overall, participants reported the arrow to be 61.25% reliable, which was not significantly different from the true value of 62.5%, t(55) = -.686, p = .495, d = -.092. We then examined whether any relationship existed between how reliable participants perceived the arrow to be and the $\beta_{Ctl} - \beta_{HPopp}$ difference in the Critical condition. No correlation was found, r(54) = .004, p = .979, suggesting that the strategy behind feature avoidance did not relate to how reliable participants believed the arrow to be.

Experiment 3 Discussion

In Experiment 3 we modified the stimulus design to try to tease apart the source of the feature avoidance effect established in the first two experiments. We tested two hypotheses; in both hypotheses, probabilistic spatial attention resulted in preferential encoding of the HP item, resulting in uncertainty about the actual target color on Critical trials where the spatial expectations were violated. However, according to Hypothesis 1, participants retained some information from the stimulus array but were unsure which of the (non-HP) colors had been in the target location, resulting in swapping the control nontarget's color (i.e., β_{Ctl} errors), whereas Hypothesis 2 posed that attention was so heavily biased to the HP location that only that HP item's color was encoded, so participants avoided that color by selecting the maximally different color on the color wheel (i.e., β_{HPopp} errors).

The Experiment 3 design allowed us to separate these two possibilities by unconfounding them in color space. We again found robust evidence of feature avoidance, with the results of multiple analyses more consistent with the first hypothesis: feature avoidance being driven by selecting one of the other colors seen in the stimulus display. However, the statistical comparisons were not significant, limiting us from drawing strong conclusions about the mechanism. Instead, we suggest that there may have been some role of both factors driving feature avoidance. At the very least, the results are not consistent with feature avoidance manifesting exclusively as 'clicking the opposite color on the color wheel', but may reflect a more complex form of strategic guessing that incorporates whatever information is accessible on a given trial.

General Discussion

Manipulating spatial attention has more far-reaching effects than simply redirecting the location of visual focus. Where and how we attend to a location can also impact the encoding and recall of features within our field of vision. Previous studies had begun to show how guiding attention by either deterministic top-down or bottom-up capture can lead to distinct perceptual and memory effects when colored items were shown and tested using a continuous response modality (Chen et al., 2019; Dowd & Golomb, 2019; Golomb, 2015; Golomb et al., 2014). The series of experiments in the current study was designed to investigate how experience-driven and probabilistic cues would impact the processing and reporting of a continuous feature such as color.

Our results revealed that both a spatial probability cue and probabilistic arrow pre-cue can lead to the same response phenomenon: feature avoidance. Both types of cues resulted in guidance of spatial attention to an expected high probability (HP) target location. On Valid trials, when the target probe was presented in the HP location, both types of cues resulted in general performance advantages, as expected based on prior reports (Geng & Behrmann, 2005; Posner, 1980). However, for Critical trials

where the target probe was presented in a different location, adjacent to the HP location, we had initially predicted this invalid attentional guidance to the HP nontarget could result in participants being prone to mistakenly reporting the color of that HP nontarget, similar to the feature swap errors found during dynamic shifts of attention (Dowd & Golomb, 2019; Golomb et al., 2014) or bottom-up capture (Chen et al., 2019). Instead, color responses were overwhelmingly more likely to reflect *avoidance* of the HP nontarget. Tellingly, participants reported relatively low confidence on these trials, suggesting that feature avoidance errors emerge when participants recognize their uncertainty regarding the target color and exhibit feature avoidance as a best guess/strategic response.

We suggest that this feature avoidance strategy emerges because probabilistic spatial attention results in the HP item being preferentially encoded into memory, leading to a reliance on incomplete or faulty memory representations for the other items when the spatial expectation is violated. But what information exactly are participants relying on during feature avoidance? In Experiments 1 and 2, feature avoidance errors manifested as responses clustered around the control nontarget color. Due to the spacing of items in color space, the control nontarget color also happened to be the color directly opposite the HP nontarget color on the color wheel. Experiment 3 was designed to unconfound these possible sources of feature avoidance. While the results did not definitively support one account over the other, they suggest that feature avoidance is not driven solely by clicking the opposite end of the color wheel. We suggest that feature avoidance may manifest in different ways, and this may partially reflect the content and quality of memory representations on a given trial. If there was a strong enough representation of the target color on a given trial, the participant will be able to successfully report that target color, even if it is not at the HP location. However, because participants use the probabilistic cue to deploy selective spatial attention to the HP location, on a substantial portion of Critical trials, they do not have a reliable representation of the target color. On these trials, they likely have a strong representation of the HP nontarget (its color *and* color-location binding) in working memory, but limited

information about the other stimuli. This limited information could range from no information about any of the other colors in the display, to partial information about the other colors, including perhaps the case where the other items' *colors* may be encoded into working memory, but with poor color-location bindings. Thus, for some trials, a failure to access any of the other items' colors may produce a form of feature avoidance where participants select the color on the color wheel maximally different than the HP nontarget. However, Experiment 3 suggests participants often do have at least partial information about the non-HP items, and on those trials may more employ a feature avoidance strategy of guessing from the remaining remembered colors. By this account, some of those guesses would be correct (target response), while others would result in swap errors with a control nontarget.

Thus, while this strategic responding is characterized by the errors it produces on certain trials, it may be more accurate to consider it a form of optimal behavior, i.e., relative to the alternative response pattern of erroneously selecting a nontarget color that had been closely attended, but is known to not have appeared at the target location. Interestingly, while we contend that feature avoidance is the result of a response strategy, conscious awareness and engagement in this tactic may not be necessary for it to emerge. The feature avoidance effect did not depend on individuals' explicit awareness of our probabilistic manipulations, raising the possibility that this strategy may ensue regardless of whether an individual realizes why they are choosing one color over the others. Whether this strategy was explicit/intentional or not remains an open question, but from the confidence reports it is clear that participants were less confident on feature avoidance trials, suggesting that the control nontarget's color was not truly believed to have been the color that appeared in the target location, but rather was selected as a form of strategic guess. This is in stark contrast to the swap errors found during stimulus-driven attentional capture (Chen et al., 2019), in which participants mistakenly reported the color of the salient distractor item with high (false) confidence they were correctly reporting the target color. Illusory conjunction errors also tend to be associated with higher confidence (Treisman & Schmidt,

1982). We thus propose that the feature avoidance effect is the manifestation of a response strategy employed when there is doubt concerning the correct target feature, as opposed to “true” swap errors of perception/memory where the participant believes the distractor color *is* the target color.

As suggested by feature-integration theory (Treisman & Gelade, 1980), attention is the “glue” that binds features together and is required for complete object processing. Feature avoidance appears to represent one of the potential consequences of realizing that glue had been used in the wrong location. The contrast between the feature binding errors observed in prior studies manipulating goal-directed shifts of attention, divided attention, and stimulus-driven attentional capture (Chen et al., 2019; Dowd & Golomb, 2019; Golomb, 2015; Golomb et al., 2014) and the feature avoidance effect discovered here illuminates the variability in how features may be encoded and remembered depending on the manner in which spatial attention is directed in the visual scene. In particular, the present study suggests that probabilistic allocation of attention can result in different impacts on feature processing than deterministic cues, even when the probabilistic cue is of a top-down nature, like our arrow cues. Instead of attention becoming tightly concentrated on the pre-cued locations, probabilistic cues may lead to a wider focus that allows more features to be encoded, albeit with the cued location recruiting most of the attentional resources. To use a gambling analogy, it is possible that probabilistic cues lead us to "bet" only a portion of our attentional resources on where our target will appear, while deterministic cues get us to go "all-in". This manner of attentional allocation has been considered by studies testing whether attention “matches” or “maximizes” its distribution of resources depending on the statistical regularities present (Jonides, 1980; Koehler & James, 2009). Rather than “maximizing” allocation to the HP location by focusing all attentional resources there, our results more closely align with a somewhat-“matching” attentional allocation, as there clearly appears to be a slightly broader distribution of attention. However, our study was not designed to obtain a direct measure of the precise extent attention is being distributed via the probabilistic cues, and therefore we cannot determine whether

attention was allocated to precisely match the statistical regularities we imposed. It is also possible that experience-driven learning plays a key role; our exploratory timecourse analyses suggested an amplification of certain effects over the course of the experiment for both spatial probability cues and the probabilistic arrow cues, though a more focused investigation of learning effects would be required to fully address this aspect.

Many models of attentional selection posit that the item or location eliciting the strongest signal will 'win the race' for our attention and be reported (Folk et al., 1992; Itti & Koch, 2000; Lee et al., 1999; Posner et al., 1980; Theeuwes, 1994). Similarly, models of visual working memory assume that the recalled item is the one that gained a strong/sufficient representation relative to the other possible memory items (Bays & Husain, 2008; Schurgin et al., 2020; Zhang & Luck, 2008). However, the results shown here serve as a reminder that the most strongly represented item may not always be the one that is chosen at the response stage. As shown previously in previous feature-reporting studies, a strongly represented nontarget may be more likely to be *selected* when attention is captured by it (Chen et al., 2019) or directed towards it in a deterministic, top-down manner (Dowd & Golomb, 2019; Golomb, 2015; Golomb et al., 2014), but *avoided* when guided towards it in a probabilistic manner, as we show here.

This differentiation is possible through the use of the continuous report paradigm that enables more sensitive estimates of feature reports. More simple measures of behavior may appear to produce similar effects across cues; for example, in visual search, faster reaction times and greater accuracy are typical characteristics of bottom-up (Chastain & Cheal, 2001; Harris et al., 2015; Yantis & Jonides, 1984), top-down (Leonard & Egeth, 2008; Posner, 1980), and experience-driven (Chun & Jiang, 1998; Geng & Behrmann, 2005; Jiang et al., 2013) cues that direct attention towards the target. The current study suggests that not only might these different types of cues elicit different characteristic patterns of feature errors when measured more sensitively, but there may be differences along other dimensions as

well; e.g., raising the possibility that probabilistic vs deterministic guidance may be a more fundamental difference than experience-driven vs goal-driven guidance. The theoretical distinctions and taxonomies between different types of attentional guidance is outside the scope of the current investigation, but it is an area of substantial recent interest, as the classic top-down / bottom-up dichotomy has been supplanted in popularity by the trichotomous branches of attentional influences (Anderson et al., 2021; Awh et al., 2012; Hutchinson & Turk-Browne, 2012; Theeuwes, 2019), with arguments it should be expanded into even more categories of guidance (Wolfe, 2021; Wolfe & Horowitz, 2017). The behavioral paradigm employed in the current study and previous work (Dowd & Golomb, 2019; Golomb et al., 2014) may offer some appeal in pursuing these questions.

Future research may seek to further understand which varieties of attentional cues can lead to feature binding errors or strategic responses. The findings thus far suggest that deterministic top-down cues and bottom-up capture both result in the former, while probabilistic cues are more likely to cause the latter. However, it is possible that other delineations may be drawn amongst the plethora of different types of cues that have been established and studied within other paradigms. Additionally, the inherent flexibility of probabilistic manipulations could raise questions regarding how the validity of probabilistic cues affects feature encoding and recall, as more trustworthy cues have been found to more strongly guide attention than those that are less reliable (Riggio & Kirsner, 1997). Investigations into whether the magnitude of feature avoidance increases along with the validity of the cue could inform us on how the variation of statistical regularities impact feature representations. Moreover, an interesting future direction to investigate is whether feature avoidance would also emerge under probabilistic retro-cues in working memory, or if it relies on selective preparatory attention prior to encoding. One speculation is that if feature avoidance was evident following a probabilistic retro-cue, it would potentially be a weaker effect, as the memory precision for retro-cued items tends to be worse than for pre-cued items (as suggested in Dube et al., 2019). Having a lower fidelity representation of the

items to begin with may lead to less precise avoidance, though it is also possible the opposite may be true if the non-HP items are more completely dropped from memory following a retro-cue (see Souza & Oberauer, 2016, for a review).

A broader related question is whether the degree of feature avoidance for a given trial or individual depends on how "focused" attention is on the HP location. While we do not have a direct measure of attentional bias in the current experiment, we considered whether memory precision (SD parameter) when the target appeared in the HP location compared to elsewhere could serve as a rough proxy for this at an individual subject level, such that individuals with larger differences in precision between Valid and Critical conditions might suggest a greater allocation of attention to the HP location. We explored whether this measure correlated with stronger feature avoidance across the participants in Experiments 1 and 2, but found only a marginal relationship (Pearson's $r = .231$, $p = .09$). Future work might utilize more direct measures for attentional allocation to investigate the relationship between the distribution of spatial attentional and the strength and/or type of feature avoidance.

While the most notable and consistent indicator of feature avoidance across the experiments conducted here was the large swap-like errors where participants selected the control nontarget's feature, there may be other aspects or variations of feature avoidance. For example, having a fair representation of the target feature in memory, but still seeking to avoid reporting the high-fidelity HP nontarget, could produce more subtle repulsion errors. Repulsion bias can be found in a variety of contexts when there is potential for interference between feature representations of multiple items. As discussed earlier, in similar paradigms where spatial attention is captured by a salient distractor, repulsion errors may reflect trials where the participant is attempting to avoid or disengage from distraction (Chen et al., 2019). When attention is intentionally divided across two locations and the two potential target colors are similar in feature space, small repulsion effects are also found, presumably as a differentiation mechanism (Golomb, 2015). Repulsion biases are also present in working memory

when representations compete with each other (Bae & Luck, 2017; Chunharas et al., 2019; Scotti et al., 2021). In the current study, we found some evidence for more subtle repulsion errors (mean shift) alongside the large nontarget feature avoidance reports. A significant shift of the target response distribution away from the HP nontarget feature (repulsion) was found in Experiment 2, and this repulsion effect was marginal in Experiment 1 (and appeared to increase over time). In Experiment 3, the asymmetric distribution of the modeled parameters prevented reliable assessment of repulsion. We also consider that feature avoidance may stem from various sources, such that it could manifest as repulsion bias, swap errors, and/or a coarse, imprecise avoidance resulting in asymmetric selection of a large section of color space; future, more targeted investigations may be better positioned to tease apart these different variations of feature avoidance and their implications.

To conclude, this study found evidence for a unique impact of probabilistic spatial cues on how stimulus features were encoded and recalled. This novel pattern of feature errors, labeled feature avoidance, manifests as a tendency to avoid an HP nontarget's feature when the correct response is unknown. In lieu of the feature binding errors elicited by other types of attentional manipulations, feature avoidance appears to be a strategic response pattern participants engage in when they know they are unsure of the target's feature. These results provide evidence for the importance of considering various aspects of how attention is guided to a spatial location, and for measures and analyses that have the precision to detect more subtle behavioral patterns in order to compare the impacts of different types of attentional guidance on feature representations.

Acknowledgements

The authors gratefully acknowledge the assistance of Anisha Babu, Alexandra Haeflein, Robert Murcko, and Veronica Olaker in the recruitment of participants and data collection. This research was supported in part by NIH grant R01-EY025648 (JG) and NSF grant BCS-1848939 (JG).

Pre-Registered Experiment	*Reported in Main Text or Supplement*	*OSF Link*	*Sample Size*	*Notes*
SPMixSwap Exp 1a	S1	https://osf.io/qtj73/	27	• Spatial probability cue • Simultaneous target design • Did not meet premise
SPMixSwap Exp 1b	S2	https://osf.io/vwxp6/	28	• Spatial probability cue • Simultaneous target design • Partially met premise
SPMixSwap Exp 2	Experiment 1	https://osf.io/cqe62/	28	• Spatial probability cue • Post-cue target design • Met premise • Feature avoidance errors
SPMixSwap Exp 3	S3	https://osf.io/4eyzq/	28	• Spatial probability cue and probabilistic arrow pre-cue • Post-cue target design • Met premise • Feature avoidance errors
SPMixSwap Exp 4	S4	https://osf.io/j7652/	17	• Spatial probability cue • Post-cue target design • Three-item array for $\beta_{\text{Ctl}}/\beta_{\text{HPopp}}$ decoupling • Did not meet premise • Terminated early due to spatial confound
SPMixSwap Exp 4b	S5	https://osf.io/wctpm/	28	• Spatial probability cue • Post-cue target design • Altered color-spacing for $\beta_{\text{Ctl}}/\beta_{\text{HPopp}}$ decoupling • Partially met premise • Feature avoidance errors
Prob Arrow Exp 1	Experiment 2	https://osf.io/347tg/	28	• Probabilistic arrow pre-cue • Post-cue target design • Met premise • Feature avoidance errors
Prob Arrow Exp 2	Experiment 3	https://osf.io/78b5s/	56	• Probabilistic arrow pre-cue • Post-cue target design • Altered color-spacing for $\beta_{\text{Ctl}}/\beta_{\text{HPopp}}$ decoupling • Met premise • Feature avoidance errors

Table 1. Reference for all pre-registered experiments in chronological order. Three experiments are included here in the main text, with results from an additional five experiments that can be found in the supplemental materials.

	Experiment 1 (blocks = 9)		**Experiment 2 (blocks = 8)**		**Experiment 3 (blocks = 8)**	
Condition	*Valid*	*Critical*	*Valid*	*Critical*	*Valid*	*Critical*
Random Guess	slope = -.0152 r = -.928 p < .001	slope = -.0104 r = -.628 p = .070	slope = -.0147 r = -.808 p = .015	slope = -.0171 r = -.639 p = .088	slope = -.0034 r = -.464 p = .246	slope = .0098 r = .604 p = .113
Standard Deviation	slope = -.2254 r = -.666 p = .050	slope = -.2600 r = -.475 p = .197	slope = -.4120 r = -.771 p = .025	slope = -.1627 r = -.210 p = .619	slope = -.1536 r = -.543 p = .164	slope = -.3936 r = -.714 p = .047
Mean Shift	slope = .1147 r = .399 p = .288	slope = -.3058 r = -.573 p = .107	slope = .1047 r = .698 p = .054	slope = -.1469 r = -.281 p = .500	slope = -.0955 r = -.397 p = .330	slope = -.0253 r = -.090 p = .831
$\beta_{CtlA/HP}$ **Response**	slope = -.0037 r = -.782 p = .013	slope = -.0039 r = -.643 p = .062	slope = -.0018 r = -.681 p = .063	slope = -.0068 r = -.892 p = .003	slope = -.0009 r = -.756 p = .030	slope = -.0044 r = -.806 p = .016
$\beta_{CtlB/Ctl}$ **Response**	slope = -.0029 r = -.739 p = .023	slope = -.0021 r = -.379 p = .315	slope = .0003 r = .169 p = .690	slope = -.0014 r = -.340 p = .410	slope = .0003 r = .437 p = .279	slope = -.0060 r = -.809 p = .015
$\beta_{CtlAopp/HPopp}$ **Response**	-	-	-	-	slope = -.0015 r = -.638 p = .088	slope = .0003 r = .053 p = .901

Table 2. Slope and correlation results for parameter estimates by block for all experiments. In order to observe any changes in our measures across time, response data was combined across participants within each experimental block (trials per "block" adjusted for Experiment 1 analyses to better match the later experiments) and then fit to the same probabilistic mixture model used for the main analyses in each experiment. The data points (one for each block) for each parameter of interest were then tested for correlations (Pearson's r) with block number, and the slope for the (linear) best-fitting line through the data points was calculated.

References

Anderson, B. A., Kim, H., Kim, A. J., Liao, M.-R., Mrkonja, L., Clement, A., & Grégoire, L. (2021). The past, present, and future of selection history. *Neuroscience & Biobehavioral Reviews*, *130*, 326–350. https://doi.org/10.1016/j.neubiorev.2021.09.004

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443. https://doi.org/10.1016/j.tics.2012.06.010

Bae, G.-Y., & Luck, S. J. (2017). Interactions between visual working memory representations. *Attention, Perception, & Psychophysics*, *79*(8), 2376–2395. https://doi.org/10.3758/s13414-017-1404-8

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7. https://doi.org/10.1167/9.10.7

Bays, P. M., & Husain, M. (2008). Dynamic Shifts of Limited Working Memory Resources in Human Vision. *Science*, *321*(5890), 851–854. https://doi.org/10.1126/science.1158023

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Chastain, G., & Cheal, M. (2001). Attentional capture with various distractor and target types. *Perception & Psychophysics*, *63*(6), 979–990. https://doi.org/10.3758/BF03194517

Chen, J., Leber, A. B., & Golomb, J. D. (2019). Attentional capture alters feature perception. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(11), 1443–1454. https://doi.org/10.1037/xhp0000681

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178. https://doi.org/10.1016/S1364-6613(00)01476-5

Chun, M. M., & Jiang, Y. (1998). Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology*, *36*(1), 28–71. https://doi.org/10.1006/cogp.1998.0681

Chunharas, C., Rademaker, R. L., Brady, T. F., & Serences, J. (2019). *Adaptive memory distortion in visual working memory* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/e3m5a

Dowd, E. W., & Golomb, J. D. (2019). Object-Feature Binding Survives Dynamic Shifts of Spatial Attention. *Psychological Science*, *30*(3), 343–361. https://doi.org/10.1177/0956797618818481

Dube, B., Lockhart, H., Rak, S., Emrich, S., & Al-Aidroos, N. (2019). *Limits to the flexible re-distribution of visual working memory resources after encoding*. PsyArXiv. https://doi.org/10.31234/osf.io/kmqtr

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 1030–1044. https://doi.org/10.1037/0096-1523.18.4.1030

Geng, J. J., & Behrmann, M. (2005). Spatial probability as an attentional cue in visual search. *Perception & Psychophysics*, *67*(7), 1252–1268. https://doi.org/10.3758/BF03193557

Golomb, J. D. (2015). Divided spatial attention and feature-mixing errors. *Attention, Perception, & Psychophysics*, *77*(8), 2562–2569. https://doi.org/10.3758/s13414-015-0951-0

Golomb, J. D., L'Heureux, Z. E., & Kanwisher, N. (2014). Feature-Binding Errors After Eye Movements and Shifts of Attention. *Psychological Science*, *25*(5), 1067–1078. https://doi.org/10.1177/0956797614522068

Harris, A. M., Becker, S. I., & Remington, R. W. (2015). Capture by colour: Evidence for dimension-specific singleton capture. *Attention, Perception, & Psychophysics*, *77*(7), 2305–2321. https://doi.org/10.3758/s13414-015-0927-0

Hutchinson, J. B., & Turk-Browne, N. B. (2012). Memory-guided attention: Control from multiple memory systems. *Trends in Cognitive Sciences*, *16*(12), 576–579. https://doi.org/10.1016/j.tics.2012.10.003

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10), 1489–1506. https://doi.org/10.1016/S0042-6989(99)00163-7

Jiang, Y. V., Sha, L. Z., & Sisk, C. A. (2018). Experience-guided attention: Uniform and implicit. *Attention, Perception, & Psychophysics*, *80*(7), 1647–1653. https://doi.org/10.3758/s13414-018-1585-9

Jiang, Y. V., Swallow, K. M., Rosenbaum, G. M., & Herzig, C. (2013). Rapid acquisition but slow extinction of an attentional bias in space. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 87–99. https://doi.org/10.1037/a0027611

Jonides, J. (1980). Towards a model of the mind's eye's movement. *Canadian Journal of Psychology / Revue Canadienne de Psychologie*, *34*, 103–112. https://doi.org/10.1037/h0081031

Kleiner, M., Brainard, D. H., & Pelli, D. (2007). *What's new in Psychtoolbox-3?* 89.

Koehler, D. J., & James, G. (2009). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition*, *113*(1), 123–127. https://doi.org/10.1016/j.cognition.2009.07.003

Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, *2*(4), Article 4. https://doi.org/10.1038/7286

Leonard, C. J., & Egeth, H. E. (2008). Attentional guidance in singleton search: An examination of top-down, bottom-up, and intertrial factors. *Visual Cognition*, *16*(8), 1078–1091. https://doi.org/10.1080/13506280701580698

Luck, S. J., Gaspelin, N., Folk, C. L., Remington, R. W., & Theeuwes, J. (2021). Progress toward resolving the attentional capture debate. *Visual Cognition*, *29*(1), 1–21. https://doi.org/10.1080/13506285.2020.1848949

Müller, H. J., & Rabbitt, P. M. (1989). Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 315–330. https://doi.org/10.1037/0096-1523.15.2.315

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). FMRI evidence for objects as the units of attentional selection. *Nature*, *401*(6753), Article 6753. https://doi.org/10.1038/44134

Olson, I. R., & Chun, M. M. (2001). Temporal contextual cuing of visual attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1299–1313. https://doi.org/10.1037/0278-7393.27.5.1299

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Posner, M. I. (1980). Orienting of Attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. https://doi.org/10.1080/00335558008248231

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*(2), 160–174. https://doi.org/10.1037/0096-3445.109.2.160

Riggio, L., & Kirsner, K. (1997). The relationship between central cues and peripheral cues in covert visual orientation. *Perception & Psychophysics*, *59*(6), 885–899. https://doi.org/10.3758/BF03205506

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), Article 11. https://doi.org/10.1038/s41562-020-00938-0

Scotti, P. S., Hong, Y., Leber, A. B., & Golomb, J. D. (2021). Visual working memory items drift apart due to active, not passive, maintenance. *Journal of Experimental Psychology: General*, *150*(12), 2506–2524. https://doi.org/10.1037/xge0000890

Sha, L. Z., Remington, R. W., & Jiang, Y. V. (2017). Short-term and long-term attentional biases to frequently encountered target features. *Attention, Perception, & Psychophysics*, *79*(5), 1311–1322. https://doi.org/10.3758/s13414-017-1317-6

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, *78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, *13*(10), 9. https://doi.org/10.1167/13.10.9

Theeuwes, J. (1994). Stimulus-driven capture and attentional set: Selective search for color and visual abrupt onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 799–806. https://doi.org/10.1037/0096-1523.20.4.799

Theeuwes, J. (2019). Goal-driven, stimulus-driven, and history-driven selection. *Current Opinion in Psychology*, *29*, 97–101. https://doi.org/10.1016/j.copsyc.2018.12.024

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*(1), 107–141. https://doi.org/10.1016/0010-0285(82)90006-8

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 11. https://doi.org/10.1167/4.12.11

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, *28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), Article 3. https://doi.org/10.1038/s41562-017-0058

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 601–621. https://doi.org/10.1037/0096-1523.10.5.601

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), Article 7192. https://doi.org/10.1038/nature06860