Template changes with perceptual learning are driven by feature informativeness

Ilmari Kurki

Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA, USA
Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA, USA
Department of Behavioral Sciences, University of Helsinki, Helsinki, Finland

Miguel P. Eckstein

Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA, USA
Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA, USA

Perceptual learning changes the way the human visual system processes stimulus information. Previous studies have shown that the human brain's weightings of visual information (the perceptual template) become better matched to the optimal weightings. However, the dynamics of the template changes are not well understood. We used the classification image method to investigate whether visual field or stimulus properties govern the dynamics of the changes in the perceptual template. A line orientation discrimination task where highly informative parts were placed in the peripheral visual field was used to test three hypotheses: (1) The template changes are determined by the visual field structure, initially covering stimulus parts closer to the fovea and expanding toward the periphery with learning; (2) the template changes are object centered, starting from the center and expanding toward edges; and (3) the template changes are determined by stimulus information, starting from the most informative parts and expanding to less informative parts. Results show that, initially, the perceptual template contained only the more peripheral, highly informative parts. Learning expanded the template to include less informative parts, resulting in an increase in sampling efficiency. A second experiment interleaved parts with high and low signal-to-noise ratios and showed that template reweighting through learning was restricted to stimulus elements that are spatially contiguous to parts with initial high template weights. The results suggest that the informativeness of features determines how the perceptual template changes with learning. Further, the template expansion is constrained by spatial proximity.

Introduction

Practice with visual tasks can lead to substantial improvements in human performance. Psychophysical studies have shown strong and persistent learning effects across a variety of tasks such as orientation and spatial frequency discrimination (Fiorentini & Nicoletta, 1980), vernier acuity (Saarinen & Levi, 1995), and texture segmentation (Ahissar & Hochstein, 1993). Many studies (Dosher & Lu, 1998; Eckstein, Abbey, Pham, & Shimozaki, 2004a; Gold, Bennett, & Sekuler, 1999; Li, Klein, & Levi, 2008; Li, Levi, & Klein, 2004) support the idea that perceptual learning improves visual processing through increasing the effectiveness of sampling of task-relevant stimulus information. More efficient sampling implies physiologically that the visual system learns to utilize neural detectors that sample more informative, high signal-to-noise parts of the stimulus and downweights detectors that sample low signal-to-noise parts of the stimulus.

Here, we use the classification image methodology (Ahumada & Lovell, 1971; Eckstein & Ahumada, 2002; Murray, 2011) and the position noise paradigm (Li et al., 2004) to estimate what stimulus parts and features the visual system samples and processes at various stages of learning. The classification image methodology uses human psychophysical decisions and external noise to estimate how the visual system weights the input information for perceptual decisions. The set of weights is referred to as the perceptual template. The

Citation: Kurki, I., & Eckstein, M. P. (2014). Template changes with perceptual learning are driven by feature informativeness. *Journal of Vision, 14*(11):6, 1–18, http://www.journalofvision.org/content/14/11/6, doi:10.1167/14.11.6.

doi: 10.1167/14.11.6 Received November 22, 2013; published September 5, 2014 ISSN 1534-7362

perceptual template is estimated from the correlation between trial-to-trial noisy stimulus values and the observer's decisions. We specifically asked whether known properties of the visual system or stimulus information could explain the dynamics of changes in the template with learning.

Here, the stimuli were noisy tilted lines, comprising checkerboard elements establishing a global tilt to the left or the right. Random positional offsets (position noise) were then added to elements' spatial location. The template presents a spatial weight function that determines how the information from low-level spatial filters is integrated for the perceptual decision (line tilted left or right). By keeping track of the signal type (left or right tilted) presented on each trial, the noise values that perturb the position of the line elements, and corresponding observer responses (tilted left or right), investigators can estimate the weight to each element (classification image) that best predicts the human behavioral responses.

Previous studies have shown that with a foveally presented stimulus, the template initially samples solely the central regions of the stimulus and thus discards large parts of potential stimulus information. With learning, the template expands toward the visual periphery, leading to more optimal sampling (Dobres & Seitz, 2010; Li et al., 2004). This suggests that perceptual template changes are constrained to initiate in areas closer to the fovea and then expand toward the more peripheral locations (retinotopic constrained expansion). However, there are alternative interpretations. A second hypothesis is that observers may initially rely on visual information at the center of the stimulus (object center expansion), which often contains the most important information. A center bias has been observed in eye movements during viewing of real scenes (Tatler, 2007). A third hypothesis is that observers' perceptual template changes initiate at stimulus parts with high inherent information and then proceed toward less informative parts (stimulus information-based expansion).

First, we asked whether the perceptual template expansion always proceeds from stimulus parts that are closer to the fovea toward peripheral locations or whether expansion initiates from the most informative parts irrespective of their spatial locations. Unlike previous studies where maximal stimulus information was presented at the fovea (Dobres & Seitz, 2010; Li et al., 2004), we used a peripheral line orientation discrimination task. The standard deviation of the position noise was constant for all elements, but because the horizontal spatial separation between elements of the two possible stimuli (orientated lines) increases proportionally to the distance from the meridian, the most informative parts of the stimulus for the orientation discrimination are at the visual periphery (Figure 1).

In experiment 1 we found that initially the template samples only the most peripheral, highly informative parts. In a second experiment we investigated spatial constraints on template expansion. In contrast to experiment 1, for experiment 2 we used a curved stimulus designed so that informativeness of an element was not tied to the distance from the fixation. Instead, in experiment 2, high and low signal-to-noise ratio (SNR) elements were spatially interleaved (Figure 5). It is known that perceptual learning can be limited to spatially adjacent locations, being much more effective, if the task has already been mastered at a neighboring area (Sigman & Gilbert, 2000). The design of experiment 2 allows the investigation of whether the expansion is determined solely by stimulus information or whether it is limited to spatially contiguous stimulus parts.

Methods

Using classification images to study perceptual learning

Classification image estimation typically requires thousands of trials. To obtain statistically reliable weights that capture perceptual learning within a single experimental session (800 trials), we used a low dimensional stimulus consisting of just 16 elements. We gathered confidence ratings in this experiment instead of more commonly used binary yes/no decisions to gain greater statistical power. Finally, we used a generalized linear model (GLM; Knoblauch & Maloney, 2008) with a more statistically accurate maximum-likelihood fitting procedure in classification image estimation. The weighted sums method (Abbey, Eckstein, & Bochud, 1999; Murray, Bennett, & Sekuler, 2002) gave very similar, but noisier, results (see Appendix A). This confirms that the results reported here are not dependent on specific statistical methods to estimate the classification images.

Observers, apparatus, and stimuli

A total of 15 volunteer observers (nine females) participated in the experiments (10 in the first study, five in the second study). The experimental procedure was in accordance with the Declaration of Helsinki and was approved by the University of California Santa Barbara's Human Subject Committee. All participants gave written informed consent.

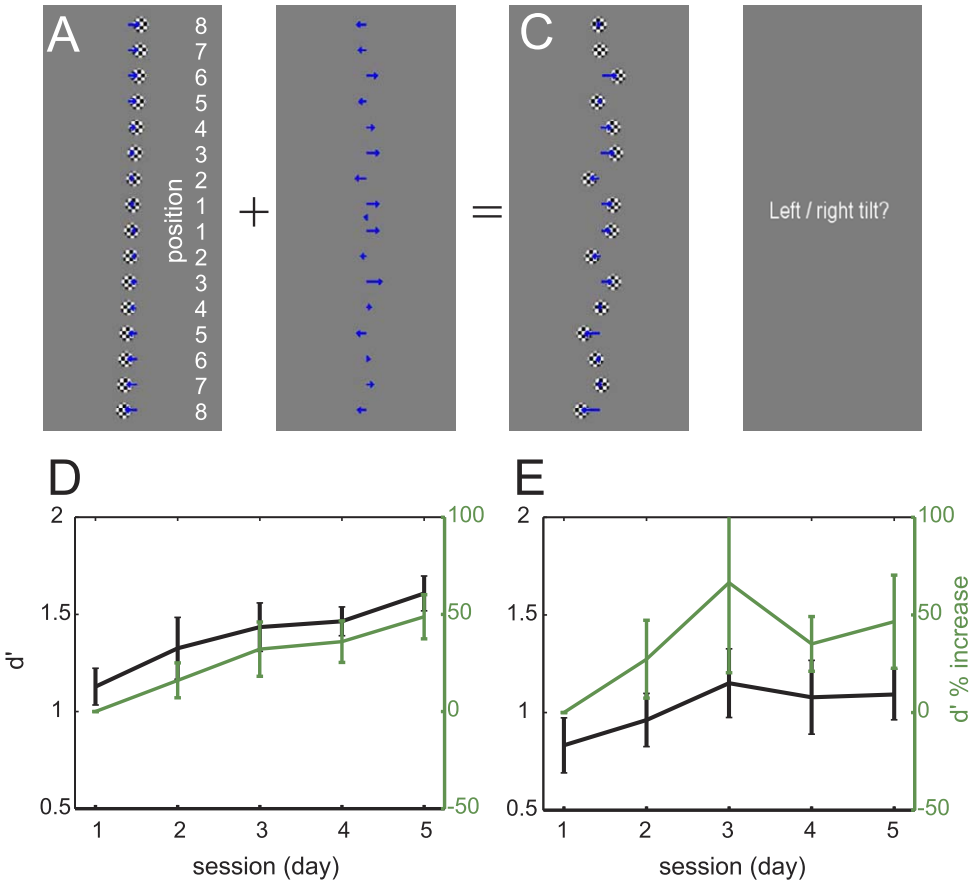

Figure 1. Stimulus in position noise paradigm and perceptual learning. A mean baseline orientation of the line was determined by sampling a leftward or rightward tilted line. Magnitude and sign of the tilt (i.e., orientation of the underlying line) were varied using the method of constant stimulus. The final stimulus presented was constructed by varying the horizontal positions of 16 elements forming the line (A). Independent random position noise values were added to stimulus elements (B) generating a noisy tilted line (C). The observer's task was to assess whether the noisy line was tilted left or right. (D) The black line represents mean performance (d′) for five sessions (10 observers) in experiment 1. The green line represents mean percentage improvement in d′ with respect to the first session. (E) Perceptual learning in experiment 2. The black line is the mean performance for five sessions (five observers), while the green line represents mean percentage improvement.

The Matlab 7.90 (Mathworks Inc., Natick, MA) environment and PsychToolbox 3.0.8 extension (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997) were used to generate and control the stimuli. An NVIDIA GEForce 7900 graphics board (NVIDIA Corp., Santa Clara, CA) hosted on a Microsoft Windows XP PC workstation (Microsoft, Redmond, WA) and a calibrated and linearized ViewSonic G90FB cathode ray tube (CRT) monitor (36 cm × 27 cm display size; 1024 × 768 pixel resolution; 100 Hz refresh rate; ViewSonic, Walnut, CA) were used to display the stimuli. Mean luminance of the display was 33 cd/m^2. The viewing distance was 69 cm. Central fixation was controlled using an Eye Link CL 1000 eye tracker (SR Research, Kanata, Ontario, Canada).

Noisy tilted lines comprised 16 circular checkerboard elements. Radius of a checkerboard element was 0.25°. The checkerboard elements were spatially placed 1° apart. The center of the line was placed 10° from the fixation at the horizontal meridian of the monitor (see Figure 1). We wanted to investigate the effects of stimulus information under conditions where visual field properties would be as constant as possible, avoiding having part of the stimulus falling on the foveal region. Previous studies suggest that changes in visual acuity are less steep in the periphery than near the fovea (Westheimer, 1982). As the presentation was quite far in the periphery, we used high-contrast checkerboard stimuli that have a broad spatial frequency spectrum and thus are less sensitive to variations in contrast sensitivity function across the visual field (De Valois & De Valois, 1988).

Lines were generated by varying the relative horizontal positions of the elements. This established a baseline tilt to left or right. In experiment 1 the target offset formed a straight line. Independent random position perturbations (position noise) sampled from a Gaussian distribution (mean = 0°, $SD = 0.5°$) were

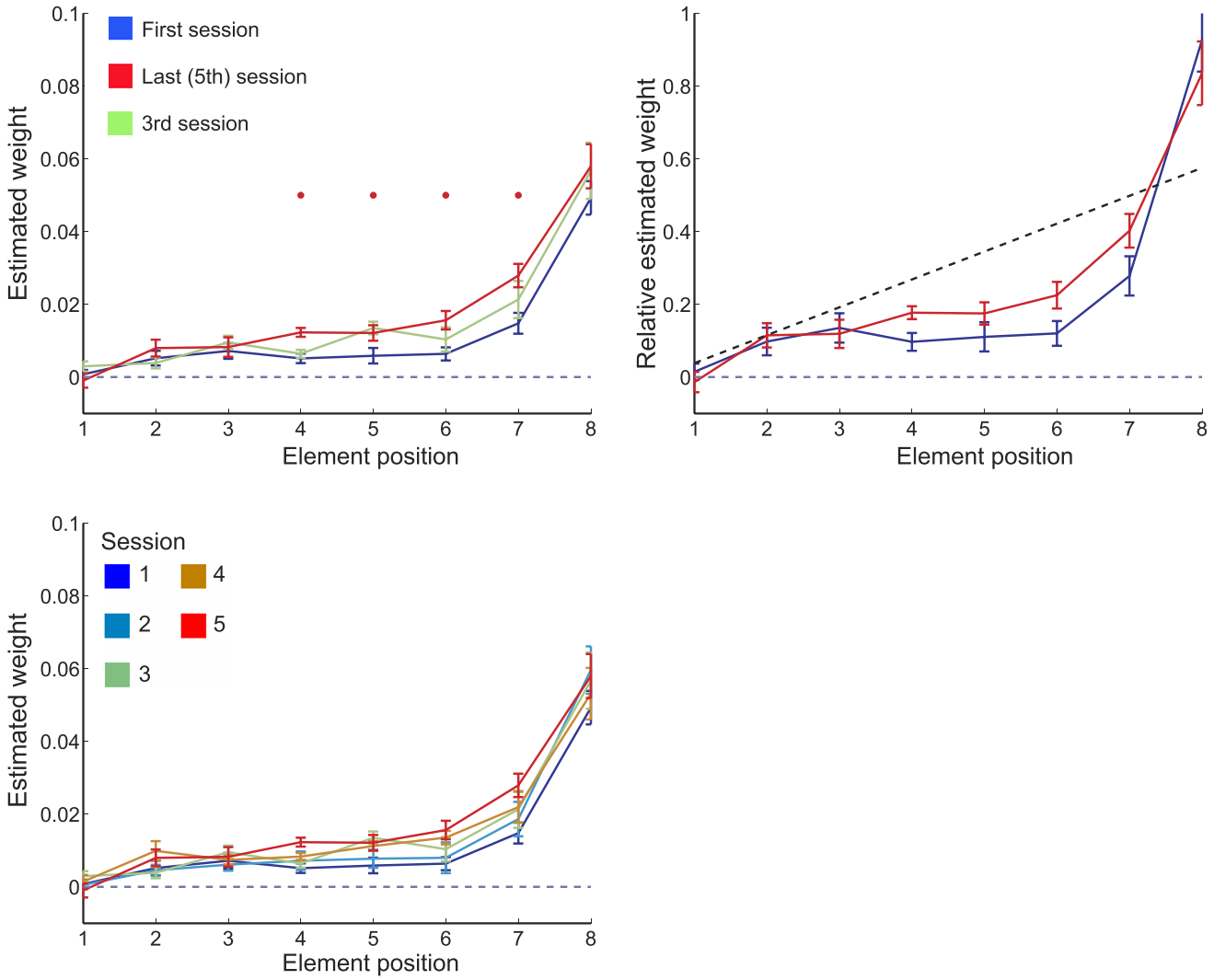

Figure 2. Over-observer classification images in experiment 1. Blue curves correspond to the first session (day), red curves correspond to the last (fifth) session, and green curves correspond to the third session. Error bars represent 1 standard error of the mean (*SEM*). Asterisks represent classification image weights with significant change between the first and last sessions ($p < 0.05$; corrected for multiple comparisons). (A) Mean classification image for eight elements for 10 observers for three sessions. (B) Normalized classification image weights for all observers. Ideal template is plotted with a dashed line. (C) Mean classification images for all five sessions.

added to the position of each checkerboard element. Position noise had a constant standard deviation at all positions, whereas the tilt offset (signal) was proportional to an element's distance from the horizontal meridian of the stimulus (see Figure 1). Therefore, the SNR or informativeness of elements grew proportional to the distance from the horizontal meridian and peaked at the element farthest from the fovea. A horizontal line was displayed as a reference in the middle of the screen to indicate to observers the location of the horizontal meridian.

In order to gain statistical power in experiment 2, we simplified the stimulus and grouped it to element pairs where neighboring elements always had the same SNR (see Figure 5). The most peripheral element group at the stimulus' ends (group *a*) had the highest SNR, as in experiment 1. However, the group next to it (group *b*; middle end) had lower SNR than the group farther toward the center of the line (group *c*; middle center). The position noise had a constant standard deviation and the same value as in experiment 1.

Procedure

The line orientation discrimination task used a method of constant stimulus (MOCS) with five levels of baseline tilt (0°–4° line tilt angle with respect to the vertical axis). Each trial started with the presentation of a fixation cue at the center of the screen for 350 ms.

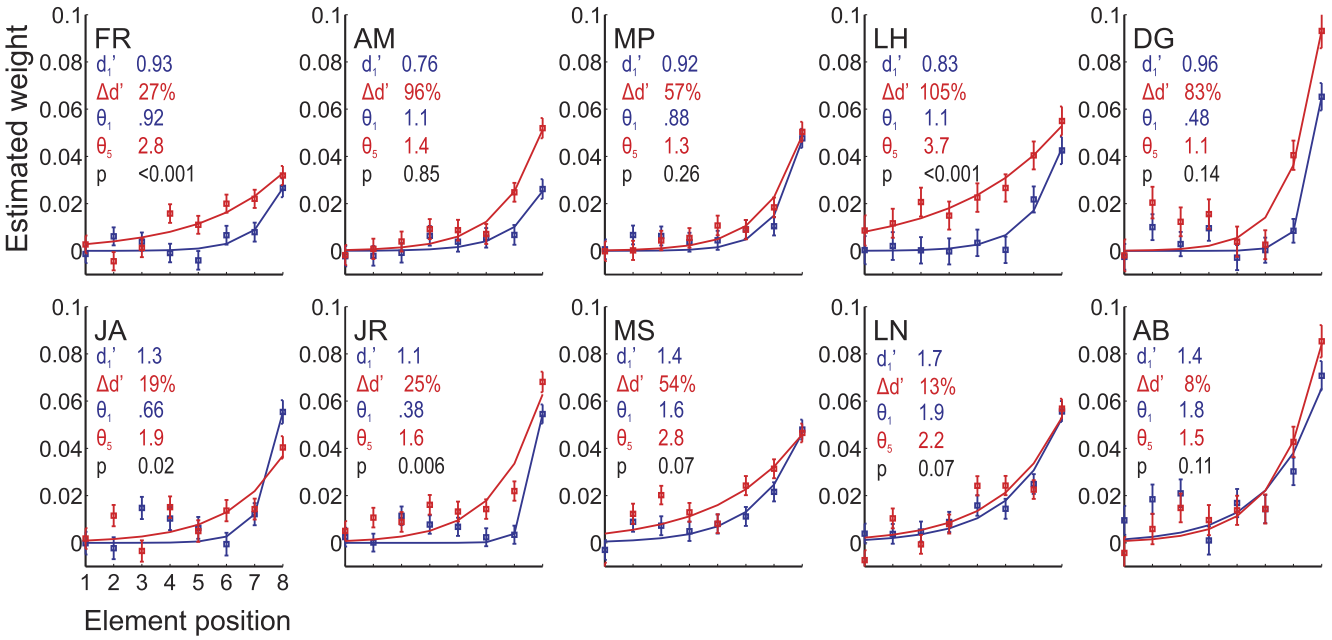

Figure 3. Individual results in experiment 1. Classification images were fitted with exponential functions (solid lines). d_1 = initial discrimination performance (d′) in the first session; $\Delta d'$ = percentage change in discrimination performance between the first and last sessions; θ_1 = width of the template fit (degrees) in the first session; θ_5 = width of the template fit in the last session; p = *p*-value of the nested likelihood test for template equality in the first and last sessions.

Following, the cue disappeared, and after a stimulus interval of 150 ms the stimulus was presented for 400 ms. An eye tracker monitored eye position, and if a fixation drift of more than 2° from the central fixation was detected the trial was classified as a broken fixation and excluded from further analysis. After stimulus presentation, the observer had to state whether the line appeared tilted to the right or to the left, using an 8-point confidence rating response (Figure 1). After the observer responded, a visual feedback icon was displayed on incorrect trials.

Each run consisted of 100 experiment trials. Each stimulus with the same position noise sample was presented twice within a run (in randomized order) to estimate the ratio of internal noise to external noise (double-pass technique). If the eye tracker calibration was poor as reported by the Eyelink calibration procedure or if more than 20 broken fixations were recorded, the run was aborted and the observers had to repeat the entire run. During each session (day) of the experiment observers participated in eight runs, resulting in a total of 800 trials. The entire experiment entailed five sessions (40 runs; 4,000 trials) conducted within 2 to 3 weeks.

Before the start of the experiments all observers were shown noiseless versions of the target stimuli to make sure they were not uncertain of the signal's shape. Before starting experiment 2 the observers were told that the target line was not straight but rather was curved.

Classification image estimation

We modeled observers' behavior using a noisy linear integrator model (Abbey et al., 1999; Ahumada, 2002; Murray et al., 2002). Only horizontal positions of the elements were varied. On any trial, k, the 16-element stimulus position vector $\boldsymbol{s}_k$ was created by multiplying (scaling) the tilt offset $\boldsymbol{t}$ with a scalar l_k (10 separate levels, determined by MOCS) determining the direction (left/right tilt) and magnitude of baseline tilt. A random position noise vector, $\boldsymbol{n}_k$, independently generated for each trial, was added to each element's position (see Figure 1):

$$\boldsymbol{s}_k = l_k\boldsymbol{t} + \boldsymbol{n}_k \quad (1)$$

The noisy linear integrator model assumes that decisions are based on the correlation between stimulus element values ($\boldsymbol{s}_k$) and a single internal template $\boldsymbol{w}$.

Random Gaussian independent internal noise, $e_k \sim N(0,\sigma_i)$, is then added to form an internal response r_k:

$$r_k = \boldsymbol{w}^T\boldsymbol{s}_k + e_k \quad (2)$$

Following the standard signal detection model (Green & Swets, 1966), we further assume that in a confidence rating experiment with m (here, eight) response options, the observer has $m - 1$ internal criteria $c = [-\infty, c_1, c_2, \ldots, c_{m-1}, \infty]$ and gives confidence rating responses number, l, when r falls between c_l and c_{l+1}. GLM is a generalization of the linear regression model to cases where the dependent variable is not

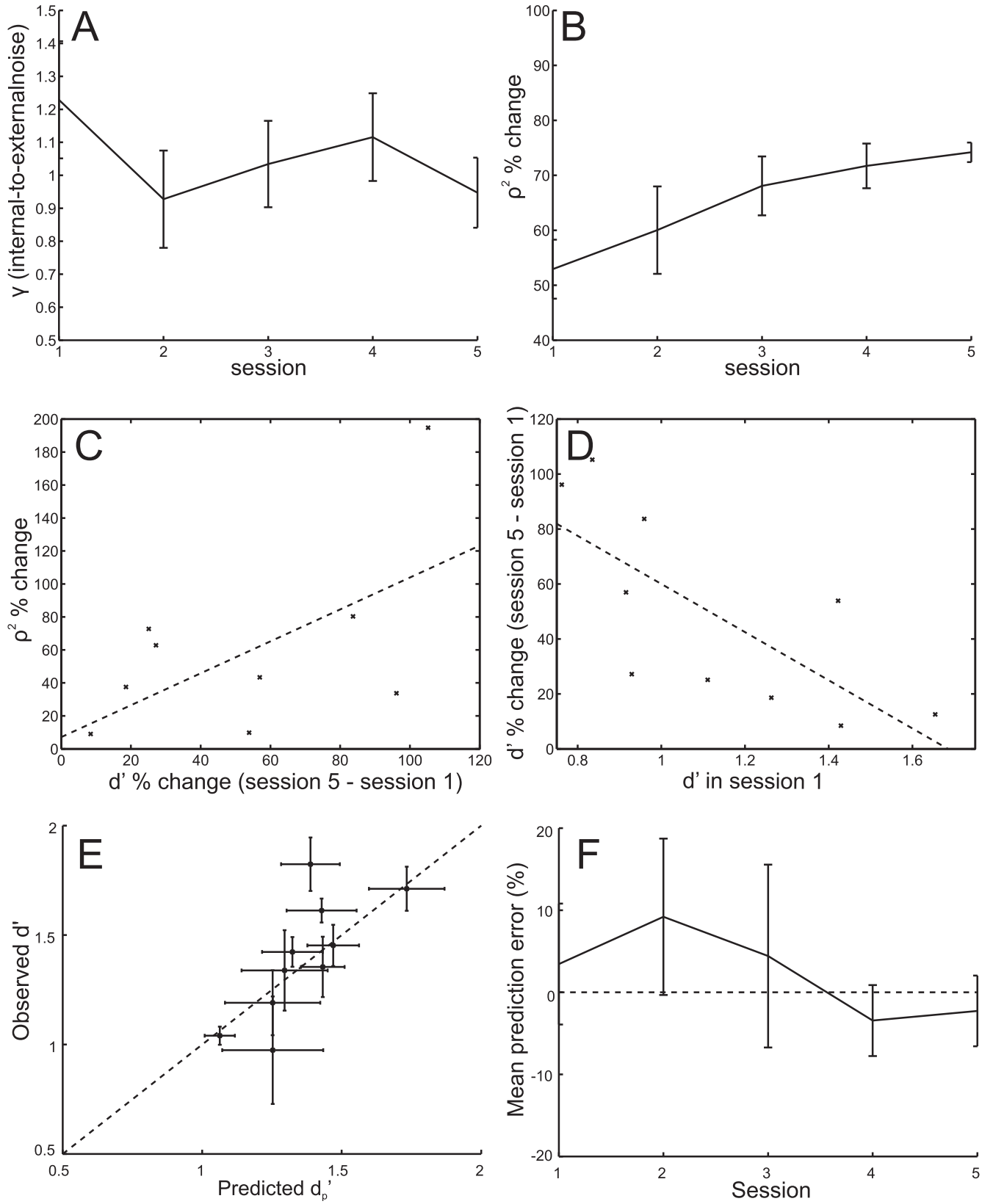

Figure 4. Comparison of template change and perceptual learning. (A) Internal-to-external noise ratio as a function of session (averaged across observers). A downward trend can be seen, but the effect was not statistically significant ($p > 0.05$). (B) Average sampling efficiency, computed by cross-correlating the estimated perceptual template and ideal template, increased significantly

→

←

across sessions (one-tailed $p = 0.007$). (C) Perceptual template efficiency is correlated with performance changes across sessions, $\rho = 0.60$, $p = 0.03$. (D) Large individual differences in both performance and amount of learning were observed. Higher performance in the first session was related to lower performance improvement across sessions, $\rho = -0.73$, $p = 0.02$. (E) Predicted d′ for the linear integrator model using estimated classification images and internal noise plotted against observed d′ for 10 observers (average of five sessions). Classification images predict observed performance, $\rho = 0.74$, $p < 0.01$, and on average, the prediction is only 2% higher than observed performance. (F) Average model prediction error for each session. Learning does not significantly increase the error ($p = 0.41$), suggesting that a linear integrator model with a template estimated from classification images and internal noise estimate can explain the majority of the observed performance and learning.

normally distributed. In GLM the dependent variable y is assumed to be a member of the exponential family of distributions (including normal, binomial, and Poisson). The expectation of observing the response y depends on a linear predictor through a possibly nonlinear link function g:

$$E(y) = g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}) \quad (3)$$

where $\boldsymbol{x}$ is vector of covariates (here, stimulus information) and $\boldsymbol{\beta}$ is a column vector of regression coefficients (template weights).

We used a GLM ordinal probit model to estimate template weights (classification images) from noise values and observers' rating scale responses using Matlab Statistics Toolbox's mnrfit function. We used only noise values as covariates while dummy variables

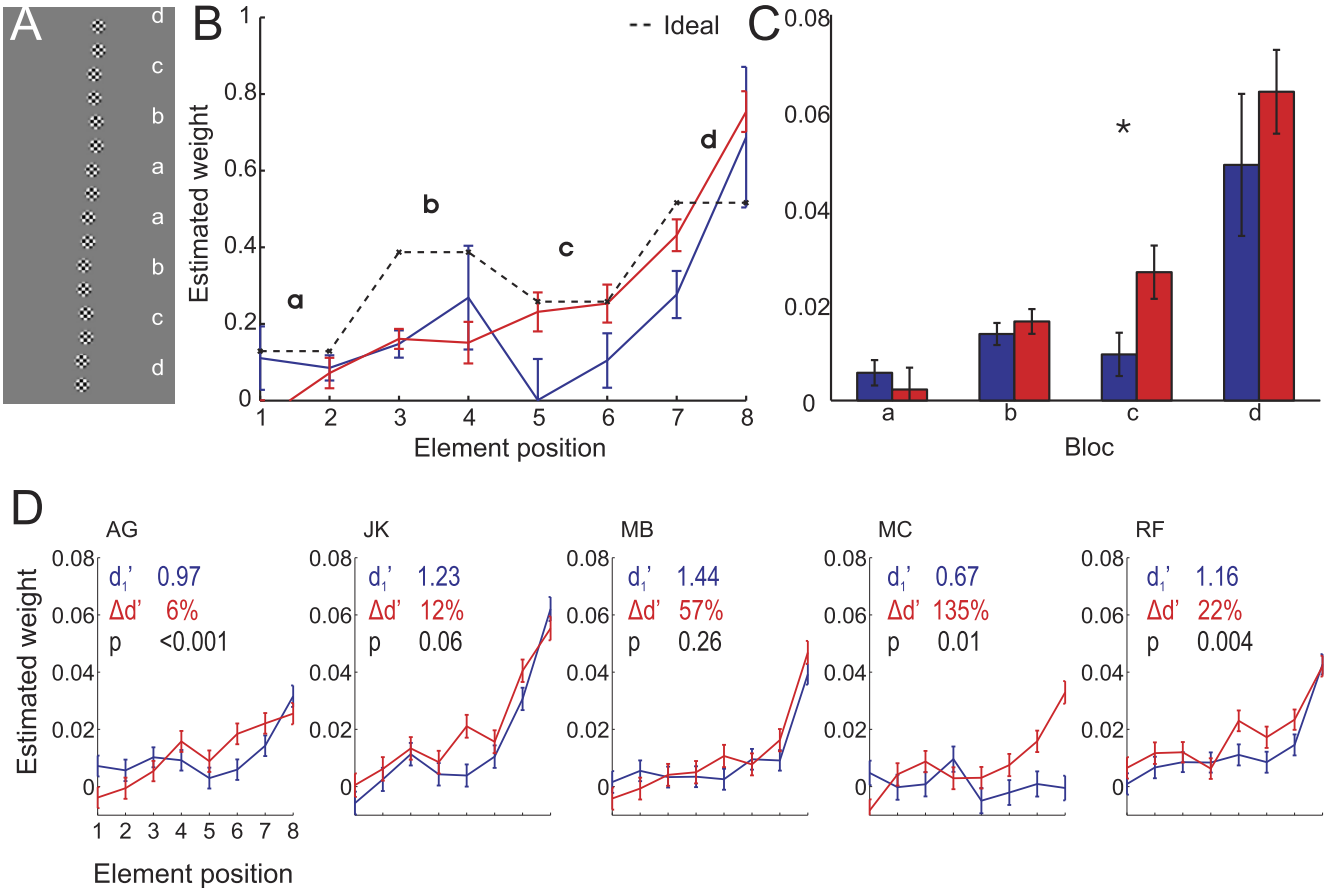

Figure 5. Stimuli and classification images in experiment 2. (A) Stimulus in experiment 2 (shown without external noise). We simplified the stimulus to four SNR groups (neighboring elements had the same SNR). Group *d* (top/bottom) had the highest SNR, group *b* the second highest, group *c* the third highest, and group *a* the lowest. (B) Average of classification images for five observers. Blue indicates the first session and red indicates the last (fifth) session; the dashed line indicates the ideal template. (C) Mean weights at different groups for the first (blue) and last (red) sessions. (D) Individual observer templates for the first (blue) and last (red) sessions. d_1 = initial discrimination performance (d′) in the first session; $\Delta d'$ = percentage change in discrimination performance between the first and last sessions; p = *p*-value of the nested likelihood test for template equality in the first and last sessions.

coded the baseline target tilt and magnitude. This ensures that any target-like patterns in the classification images could not be caused by the baseline target offset but instead reflect observers' perceptual templates (Knoblauch & Maloney, 2008; Murray, 2011). The full GLM model had four factors for internal criteria, five factors for the baseline target orientation, and 16 covariates for classification image weights.

The data were preprocessed by removing trials where observers broke fixation. We reduced the number of rating scale response categories from eight to four by merging two neighboring rating scale responses into a single category in order to ensure that each category had enough samples for the analysis. Classification images were then estimated for each session and fitted with an exponential function. In the main analysis, we analyzed the classification images by averaging the two mirror-symmetric parts of the line in order to gain statistical power.

Individual and group analyses

It is well known that there are large individual differences in perceptual learning (Beard, Levi, & Reich, 1995). Therefore, we show both individual data and average perceptual templates across observers. However, averaging classification images can be problematic as observers often have differences in both the perceptual template profile and amplitude, which is inversely related to internal noise level (Ahumada, 2002; Eckstein, Pham, & Shimozaki, 2004b). In principle, averaging of classification images might cause the resulting average to be dominated by those observers who have large absolute classification image magnitudes. To ensure that changes in the average classification images did not solely reflect changes in observers with high-amplitude classification images, we also analyzed the data by normalizing the individual classification image amplitudes by the maximum weight in the first session (not shown). The results were very similar to the simple mean classification images which are presented here.

Last, we tested the learning-related template changes at the individual-subject level using nested hypothesis likelihood ratio tests (Knoblauch & Maloney, 2008, 2012). We compared the likelihood of a model where a single classification image was estimated on both the first and the last sessions to the likelihood of a model that had two separate classification images for both sessions. Likelihood ratio test using χ^2 test statistics can be used to reject the hypothesis that both sessions had the same classification image.

Performance measures

In addition to classification images, discrimination performance was characterized by computing the index of detectability, d′, for each signal strength level of the MOCS and taking the mean across three medium signal strength levels. The lowest and highest levels were discarded. The lowest level did not have any physical signal and performance was perfect for some observers at the highest level, making d′ estimation not possible.

Sampling efficiency estimation

When independent external noise is added to each element of a stimulus, an optimal Bayesian observer weights each element proportionally to its signal amplitude and inversely to its noise variance (Burgess, Wagner, Jennings, & Barlow, 1981; Eckstein, Whiting, & Thomas, 1996; Geisler, 2011; Gold et al., 1999; Green & Swets, 1966; Peterson, Birdsall, & Fox, 1954; Tjan & Legge, 1998), which here increases with the distance from the fovea. We investigated whether observers optimized their perceptual template by improving its match to that of an optimal observer. To compare human and ideal templates we computed the sampling efficiency of human templates (ρ^2) by correlating the estimated human template, $\hat{\boldsymbol{w}}_e$, with the ideal template, $\hat{\boldsymbol{w}}_i$. The classification image without averaging over mirror symmetry was used:

$$\rho^2 = (\hat{\boldsymbol{w}}_e^{\boldsymbol{T}} \hat{\boldsymbol{w}}_i)^2 \quad (4)$$

This method does not require any assumptions of the shape of the observer template but gives a conservative estimate of true sampling efficiency, as each point of the classification images contains some estimation noise. However, with this low dimensional stimulus we got reasonably high sampling efficiency estimates, suggesting that estimation noise was not a major factor.

Internal noise reduction

Another possible mechanism by which perceptual learning might increase performance is a reduction of the amplitude (standard deviation) of the internal noise added to the observer's internal response to the stimulus' elements. To assess the role of internal noise in the learning in our task, we used the double-pass technique (Burgess & Colborne, 1988) to estimate the amount of effective internal noise on each session. Each noise (plus stimulus) sample in the experiment was shown twice within the same session. As the external noise samples across pairs of trials are exactly the same, any discrepancies between responses in two passes reflect stimulus-independent internal noise in the system. The degree of internal noise can then be estimated from average response consistency with minimal assumptions on how the observer processes

the stimuli (i.e., linearity of response; see also Ahumada, 2002). We generalized the double-pass method (Green, 1964; Burgess & Colborne, 1988) for a yes/no task with binary decisions to a yes/no task with multiple confident ratings (see also Ahumada, 2002).

Let $r_s = r_e + r_t$ be the internal response of the observer model without internal noise. Let the variance of external noise response r_e be 1, r_t be the response to the noiseless target, and $r_i \sim N(0,\sigma_i)$ be the trial-to-trial response fluctuation caused by the internal noise. We can thus define an internal-to-external noise ratio $\gamma = \sigma_i$. The probability of an observer giving response confidence ratings k and l to the same external noise values shown on two different trials given the criteria c_k, c_{k+1} is

$$p(a_1 = k, a_2 = l|r_s) = p(c_k < r_e + r_{i1} < c_{k+1})p(c_l < r_e + r_{i2} < c_{l+1}) \quad (5)$$

$$= \left(\Phi\left(\frac{c_{k+1} - r_e}{\sigma_i}\right) - \Phi\left(\frac{c_k - r_e}{\sigma_i}\right)\right) \times \left(\Phi\left(\frac{c_{l+1} - r_e}{\sigma_i}\right) - \Phi\left(\frac{c_l - r_e}{\sigma_i}\right)\right) \quad (6)$$

where $\Phi(x)$ is the standard cumulative normal distribution. We analyzed each signal strength level and target orientation separately; thus, the target response r_s can be assumed to be constant on each condition. Integrating over all external noise values gives the expectation

$$p(a_1 = k, a_2 = l) = \int_{-\infty}^{\infty} p(a_1 = k, a_2 = l|r_e)p(r_e)dr_e \quad (7)$$

$$= \int_{-\infty}^{\infty} p(a_1 = k, a_2 = l|r_e)\theta(r_e)dr_e \quad (8)$$

$$= \int_{-\infty}^{\infty} \left(\Phi\left(\frac{c_{k+1} - (r_t + r_e)}{\sigma_i}\right) - \Phi\left(\frac{c_k - (r_t + r_e)}{\sigma_i}\right)\right) \times \left(\Phi\left(\frac{c_{l+1} - (r_t + r_e)}{\sigma_i}\right) - \Phi\left(\frac{c_l - (r_t + r_e)}{\sigma_i}\right)\right) \theta(r_t + r_e)dr_n \quad (9)$$

where $\theta(x)$ is the standard normal distribution. Criteria c as well as target response r_t were estimated from the experiment's rating responses using the standard signal detection model (Green & Swets, 1966; Macmillan & Douglas, 2005). For example, r_t relates to detectability index d' simply as $r_t = \sqrt{(1+\gamma^2)}d'$ for target tilted left trials and 0 for target tilted right trials. We evaluated the expected response probabilities for stimulus pairs for different internal-to-external noise levels using an exhaustive parameter search, calculating the expected response distribution for each γ and then evaluating the best fit, by comparing the error sum of squares between the estimated and expected response distributions.

Numerical integration (quad function of Matlab) was used to solve Equation 9. Each target orientation (level) was analyzed separately for each session and the mean of these was used as an estimate of each session's internal-to-external noise ratio.

We ran simulations that showed that using more than two confidence rating responses yields more accurate internal-to-external noise estimates. However, we chose to limit the number of rating categories to four (by merging rating responses) in order to get better probability estimates from the limited number of trials in each session.

Estimation of human performance using linear model observer evaluation

Finally, we test the observer model (Equation 2) by comparing how well the noisy linear integrator model can predict the observed detection performance. It can be shown that the model's predicted detection performance (d'_p) is completely determined by specifying the template sampling efficiency (ρ^2), the energy difference between the two target signals (E_Δ), and variances of external (σ_e^2) and internal (σ_i^2) noise (Equation 10) (Burgess et al., 1981; Eckstein et al., 2004b):

$$d'_p = \sqrt{\frac{\rho^2 E_d}{\sigma_e^2 + \sigma_i^2}} \quad (10)$$

We used the sampling efficiency estimated from classification images and internal noise from the double-pass method to test how accurately the model predicts the observed human performance for our stimuli.

Results

Experiment 1: Perceptual template changes are determined by stimulus information, not visual field structure

Discrimination performance and learning differences

Figure 1D shows average discrimination performance as well as improvement relative to the first session. Detectability increased significantly across sessions, $t(9) = 5.20$, one-tailed $p = 0.0028$. A one-tailed test for d′ change was used because we tested the specific directional hypothesis that performance increased after learning. On average, the mean d' was 48% higher on the fifth session. In addition, we found

large individual differences in learning (see individual performance data in Figure 3). The learning improvement ranged from 8% to 102%. A major covariate with the differences in the magnitude of learning was an observer's performance in the first session (Figure 4D). Initial performance and subsequent improvement showed a strong negative correlation, $\rho = -0.73$, $p = 0.03$: Observers that had strong initial performance improved less than those that had weak initial performance. Similarly, template sampling efficiency on the first session was negatively correlated with performance improvement, $\rho = -0.60$, $p = 0.03$: Observers with a perceptual template that least matched with the optimal template in the first session showed the most improvement in perceptual performance in subsequent sessions.

We computed the number of times the experiment was restarted because of eye tracker calibration problems. On average there were 7.6 restarts across all 40 runs. This was not correlated with the amount of improvement in d′ across the sessions, $\rho = 0.09$, $p = 0.8$. We also computed the number of broken fixations for each observer to make sure that there was not a large variation across observers in the stimulus views. On average 5% (minimum: 2%, maximum: 12%) of the trials were discarded because of broken fixations.

Classification images

Figure 2 shows classification images for the first (blue) and last (red) learning sessions. The y-axis corresponds to the estimated relative weights an observer assigned to the stimulus elements. The x-axis corresponds to the elements' position away from the horizontal meridian. Bottom rows show individual profiles, fitted with an exponential function. The top row shows the grand average classification images, averaged across individuals. For the grand average classification images, we show both average decision weights (A) and normalized decision weights for comparison with the ideal observer (B). Panel C plots classification images for all five sessions and shows the progressive change of the perceptual template across sessions. All observers initially relied highly (in some cases, such as observers AM and JR, almost solely) on the most informative elements (highest SNR; SNR hot spot), which are the farthest from the meridian (top and bottom elements). Learning changed the perceptual template to also include elements distant from the top and bottom elements. Width of the fitted template grew on average 124%, $t(9) = 3.45$, $p = 0.007$. Amplitudes of the fits did not show systematic changes, $t(9) = -0.561$, $p = 0.59$. Average classification images show that observers' perceptual templates changed across sessions, expanding from the initial SNR hot spot toward the central meridian. Hotelling's T^2 test, a multivariate generalization of the univariate Student's t, showed that the template change was significant, $T^2(8, 9) = 3490$, $p = 0.01$. About five out of 10 observers showed significant ($p < 0.05$) template changes at the individual level. Multiple univariate two-tailed t-tests corrected for multiple comparisons using the false discovery rate correction (Benjamini & Hochberg, 1995) at $\alpha = 0.05$ show that, across sessions, the perceptual weights changed at locations close to the initial SNR hot spot (locations 4–7) while no significant change in weighting was found for the top and bottom element (location 8).

We further analyzed any potential effects of baseline tilt (signal strength) on the classification images (see Appendix B). No significant differences were found when using a false discovery rate of $\alpha = 0.05$. Because the weights for the two mirror-symmetric parts of the stimulus (bottom and top halves) did not differ significantly, $T^2(8, 9) = 14.29$, $p = 0.845$, they were averaged. Last, we compared classification images that were analyzed from trials with line baseline tilt toward the right versus left and found no significant differences due to left/right baseline tilt, $T^2(8, 9) = 65.54$, $p = 0.4$, or toward center and periphery, $T^2(8, 9) = 384.45$, $p = 0.09$ (see Appendix B).

Sampling efficiency and internal noise

Figure 4B shows that the sampling efficiency of the perceptual templates improved across sessions by 54% on average, from $\rho^2 = 0.52$ to $\rho^2 = 0.74$, $t(9) = 3.04$, one-tailed $p = 0.007$. Comparison of estimated templates and ideal templates is also shown in Figure 2B. We assessed whether observers with greater increases in sampling efficiency were associated with larger gains in perceptual performance. We found a positive correlation of $\rho = 0.60$ between template sampling efficiency improvement and performance improvement, one-tailed $p = 0.03$ (Figure 4C); thus, observers with greater template optimization showed larger improvement in the orientation discrimination task. One-tailed tests were used because we tested the specific directional hypothesis that sampling efficiency increased as a result of template change and that the correlation with sampling efficiency and performance improvement was positive.

We tested whether learning improves the internal-to-external noise ratio. If the internal noise decreases with practice, then each observer's ratings should show larger agreement when viewing the same image in the latter sessions. Figure 4A shows that there is a downward trend in the internal-to-external noise ratio, but this reduction did not reach statistical significance, $t(9) = -1.76$, one-tailed $p = 0.06$. Furthermore, unlike the sampling efficiency, the variation across observers in internal noise reduction through learning was not

significantly correlated with differences in improvements in d′, $\rho = -0.2141$, $p = 0.55$.

We found that the linear noisy integrator observer model (Equation 10) was able to predict quite accurately the observed d′ (Figure 4E; points represent average prediction across five experiment sessions for each subject). On average, model prediction (d'_p) was only 2% higher ($SE \pm 3\%$) than the observed d′, and the difference was not statistically significantly different from zero, $t(9) = -0.280$, $p = 0.98$. Correlation between the observed and the predicted d′ was high, $\rho = 0.74$, one-tailed $p = 0.014$. Figure 4F shows the average prediction error for each session. There was no statistically significant difference in prediction error between the first and last sessions, $t(9) = -0.09$, $p = 0.41$.

Experiment 2: The role of spatial contiguity in perceptual learning

Performance

Orientation discrimination performance (d') in experiment 2 improved by an average of 47% (see Figure 1E) over five sessions (maximum: 135%; minimum: 6%). The improvement in d' was significant, $t(4) = 3.08$, one-tailed $p = 0.019$.

Classification images

Experiment 2 spatially interleaved highly informative and less informative groups of elements. If learning is mostly determined by the inherent informativeness of an element, we would expect the template weighting to initiate from the highest SNR elements at the edge of the line and then progress to the second highest SNR elements at the middle center (group *c*; see Figure 5) rather than the middle end (group *b*). However, if learning is constrained by the spatial proximity to learned elements, we would expect to find the largest template weighting changes at the locations (middle end) spatially contiguous to the highest SNR element.

Figure 5C shows classification images for five observers for the first (blue) and last (red) sessions as well as the ideal perceptual template—Observers initially (first session) relied mostly on the information provided by high SNR elements (group *d*; top and bottom elements). Comparison across sessions (session 5 – session 1) reveals that practice did not change weights at the initial hot spot (high SNR) group of elements *(d)*, $t(4) = 1.78$, $p = 0.15$. However, a significant change (179% increase) was found at the location adjacent to it (group *c*), $t(4) = 6.07$, $p = 0.004$, whereas change at the group where SNR was second highest (group *b*; 19% increase) was not statistically significant, $t(4) = 2.78$, $p > 0.05$, when corrected for multiple comparisons at a false discovery rate of $\alpha = 0.05$.

Discussion

Perceptual learning: Sampling efficiency, performance, and internal noise

Both discrimination performance and sampling efficiency of the estimated perceptual templates increased with practice. Results show that a noisy linear integrator model can predict rather accurately (2% prediction error) the observed detection performance. Moreover, the accuracy of the linear model's prediction of human performance was similar in the first and last sessions. Thus, template changes and internal noise reduction can explain to a great extent the observed learning, and other potential factors that are not captured by the linear integrator model, such as nonlinearities (Lu & Dosher, 2008; Zhang, Pham, & Eckstein, 2006), might play only a marginal role for the current task.

We found a positive correlation between increases in sampling efficiency and performance improvements, suggesting that the template changes can account for the behavioral learning. These results provide further evidence that perceptual learning operates mainly by optimizing the perceptual template (i.e., the visual system's sampling and integration of the stimulus information) (Abbey, Pham, Shimozaki, & Eckstein, 2008; Dosher & Lu, 1998; Eckstein et al., 2004a; Gold et al., 1999; Gold, Sekuler, & Bennett, 2004; Li et al., 2004; Peterson, Abbey, & Eckstein, 2009; Trenti, Barraza, & Eckstein, 2010). Performance and sampling efficiency showed the largest improvements in early learning sessions, in agreement with many previous studies (see, e.g., Gold et al., 2004; Karni & Sagi, 1993).

On the other hand, using the double-pass method, arguably the most reliable way to estimate the amount of internal noise (Ahumada, 2002), we did not find a reliable reduction in internal noise. Furthermore, observer variations across sessions in the internal noise did not correlate with observer differences in performance improvement. This finding is in agreement with previous evidence on perceptual learning, suggesting that reduction of stimulus-independent internal noise compared with template changes has a smaller role in perceptual learning for a variety of tasks (Gold et al., 1999, 2004; Li et al., 2004).

We found large individual differences in the magnitude of the performance increases (d' improvement ranging from 8% to 102%). Interestingly, the magnitude of observers' performance improvement was inversely correlated with observers' initial performance

(Figure 4D): Observers with low initial performance showed more learning. Similar results were reported by Dobers and Seitz (2010). Also, Li et al. (2008) reported a similar effect in amblyopic observers, whose initial visual acuity in a vernier task was inversely related to the magnitude of improvement and time-constant of learning. On the other hand, low initial sampling efficiency in observers was positively correlated with performance improvement. This further demonstrates that a major part of perceptual performance improvements can be explained by observers' template changes: Observers with the least optimal initial templates show the most performance improvement. It is likely that observers with high sampling efficiency templates underwent rapid learning within the first session, precluding reliable estimation of the changes in their perceptual templates. For most observers, improvements in both performance and sampling efficiency saturated after the second session, with a trend of modest improvement in subsequent sessions.

Changes in perceptual learning: Foveal to periphery, stimulus center, or stimulus information-based expansion

The main goal of the study was to assess how learning changes the sampling of the stimulus parts and features in a line orientation task and to identify what factors determine the dynamics of the perceptual template changes (expansion). We assessed three hypotheses: (1) The structure of the visual field results in perceptual templates, with initial weighting of stimulus elements at the fovea and a subsequent expansion toward the visual periphery (retinotopic constrained expansion); (2) the template expands from the center of the stimulus toward the outer parts of the stimulus (object-centered expansion); and (3) the perceptual template changes are determined by the informativeness of stimulus parts and features. Templates initially weight only highly informative elements and progress toward elements with lower information content (stimulus information-based expansion).

The results show that the template expansion is not constrained to be from the fovea to the visual periphery (foveal to periphery expansion) or from the center of an object toward its outer boundaries (stimulus center expansion). Instead, the template expansion is determined by element or stimulus part informativeness, with initial weightings given to the elements with the highest SNR (stimulus information-based expansion).

Li et al. (2004) and Dobres and Seitz (2010) demonstrated an expansion of the perceptual template from the fovea to the periphery. However, results from these two studies can also be explained by stimulus information-based expansion. Although the information content in their stimuli was kept constant across the visual field, the inhomogeneous processing across the retina might result in elements at the center of the stimulus, which were processed with the high-resolution fovea (De Valois & De Valois, 1988), to have higher quality information. In the present study the more peripheral elements contained more information than did the less peripheral elements of the stimulus. This experimental design allowed us to distinguish between the various hypotheses for template expansion. Our results clarify the factors governing the dynamics in template tuning during perceptual learning. Initially, observers rely on local, high signal-to-noise parts of the stimulus. In our study these corresponded to more peripheral elements. In many common scenarios for which the stimulus has homogeneous inherent information, the parts with highest effective information content might correspond to stimulus parts processed with the high-resolution fovea (Dobres & Seitz, 2010; Li et al., 2004). Without learning, the ability to integrate larger areas of stimulus is limited. During learning, the perceptual templates expand to cover larger areas of the stimuli, increasing the efficiency of sampling. However, previous studies show that template optimization does not necessarily always involve template expansion because some stimuli result in initial broad templates that, with learning, shrink to be better tuned to the information in the stimulus (Kurki, Hyvärinen, & Laurinen, 2006).

In the second experiment we found that when high and low SNR elements were interleaved, observers initially relied mostly on the high signal-to-noise elements and, subsequently, practice changed template weights only on locations spatially close to parts that were represented in the initial template. Weights at more informative (higher SNR) elements farther away were virtually unchanged. Thus, improved processing caused by learning tends to spread retinotopically across the visual field from already-sampled locations to spatially contiguous locations. Perceptual learning studies in visual search (Sigman & Gilbert, 2000) and pop-out detection (Ahissar & Hochstein, 2000) have also reported retinotopic spreading of learning. However, in these studies the retinotopic spreading does not have any immediate cost and might be a good adaptive strategy when there is some spatial uncertainty in the retinotopic location of the stimulus. In contrast, in experiment 2 of the current study, retinotopic spreading of the perceptual template is a suboptimal strategy: Increasing the perceptual template weight at the second most informative location (group *b*) would improve perceptual performance, while increasing the weight of a low-SNR element (group *c*) contiguous to the highest SNR element would contribute less to performance improvements. Thus, our results show that perceptual learning is constrained by retinotopic proximity of

informative stimulus elements. Some perceptual learning models have postulated mechanisms that could explain such constraint. One popular idea is that perceptual learning operates by changing the local connectivity in early visual areas (see, e.g., Adini, Sagi, & Tsodyks, 2002; Sigman & Gilbert, 2000). This account would explain why perceptual template change here was constrained to spatially close areas.

On the other hand, spatial attention is assumed to determine what information can be learned (e.g., Ahissar & Hochstein, 1993; Fahle, 2004; however, see Watanabe, Náñez, & Sasaki, 2001). Many models of attention assume spatially restricted local spotlight of attentional control in the visual field. Under these assumptions learning would be spatially restricted to neighboring areas as the attentional system is not able to distribute control arbitrarily, even when updating template weights at a more distant location would gain more benefit in the task. Thus, retinotopic constraints per se do not imply that learning occurs at early visual areas (Mollon & Danilova, 1996; Petrov, Dosher, & Lu, 2005; Xiao, Zhang, Wang, & Klein, 2008).

Conclusions

We used the classification image technique and showed how observers changed their templates in a discrimination of a peripherally located orientated lines with elements that varied in SNR. We tested various hypotheses about the factors influencing perceptual template changes: template expansion driven by the retinotopic structure of the visual field, template expansion governed by the center of the object (object centered), and template expansion determined by the informativeness of the stimulus parts. Our results favor the hypothesis that changes in the perceptual template progress from the most informative parts and features to those that are less informative. However, we show that template changes are constrained to spatially contiguous regions of the stimulus. Together, the results increase our understanding of the influences and neurobiological constraints of human perceptual learning.

Keywords: perceptual learning, classification image, psychophysics

Acknowledgments

Supported by grant EY-015925 from the National Institutes of Health, National Eye Institute (ME) and grants from Oskar Huttunen Foundation (IK) and ASLA—Fulbright (IK). The authors declare no competing financial interests.

Commercial relationships: none.
Corresponding author: Ilmari Kurki.
Email: ilmari.kurki@helsinki.fi.
Address: Department of Behavioral Sciences, University of Helsinki, Helsinki, Finland.

References

Abbey, C. K., Eckstein, M. P., & Bochud, F. O. (1999). Estimation of human-observer templates in two-alternative forced-choice experiments. *Proceedings of SPIE, 3663,* 284–295.

Abbey, C. K., Pham, B. T., Shimozaki, S. S., & Eckstein, M. P. (2008). Contrast and stimulus information effects in rapid learning of a visual task. *Journal of Vision, 8*(2):8, 1–14, http://www.journalofvision.org/content/8/2/8, doi:10.1167/8.2.8. [PubMed] [Article]

Adini, Y., Sagi, D., & Tsodyks, M. (2002). Context-enabled learning in the human visual system. *Nature, 415,* 790–793.

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences, USA, 90,* 5718–5722.

Ahissar, M., & Hochstein, S. (2000). The spread of attention and learning in feature search: Effects of target distribution and task difficulty. *Vision Research, 40,* 1349–1364.

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision, 2*(1):8, 121–131, http://www.journalofvision.org/content/2/1/8, doi:10.1167/2.1.8. [PubMed] [Article]

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America, 49,* 1751–1756.

Beard, B. L., Levi, D. M., & Reich, L. N. (1995). Perceptual learning in parafoveal vision. *Vision Research, 35,* 1679–1690.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57,* 289–300.

Brainard, D. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436.

Burgess, A., & Colborne, B. (1988). Visual signal

detection. IV. Observer inconsistency. *Journal of the Optical Society of America A, 5,* 617–627.

Burgess, A., Wagner, R., Jennings, R., & Barlow, H. (1981). Efficiency of human visual signal discrimination. *Science, 214,* 93–94.

De Valois, R. L., & De Valois, K. K. (1988). *Spatial vision*. New York: Oxford University Press.

Dobres, J., & Seitz, A. R. (2010). Perceptual learning of oriented gratings as revealed by classification images. *Journal of Vision, 10*(13):8, 8–11, http://www.journalofvision.org/content/10/13/8, doi:10.1167/10.13.8. [PubMed] [Article]

Dosher, B. A., & Lu, Z. -L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences, USA, 95,* 13988–13993.

Eckstein, M., Abbey, C. K., Pham, B. T., & Shimozaki, S. S. (2004a). Perceptual learning through optimization of attentional weighting: Human versus optimal Bayesian learner. *Journal of Vision, 4*(12):3, 1006–1019, http://www.journalofvision.org/content/4/12/3, doi:10.1167/4.12.3. [PubMed] [Article]

Eckstein, M., & Ahumada, A. (2002). Classification images: A tool to analyze visual strategies. *Journal of Vision, 2*(1):i, http://www.journalofvision.org/content/2/1/i, doi:10.1167/2.1.i. [PubMed] [Article]

Eckstein, M., Whiting, J. S., & Thomas, J. P. (1996). Role of knowledge in human visual temporal integration in spatiotemporal noise. *Journal of the Optical Society of America A, 13,* 1960–1968.

Eckstein, M. P., Pham, B. T., & Shimozaki, S. S. (2004b). The footprints of visual attention during search with 100% valid and 100% invalid cues. *Vision Research, 44,* 1193–1207.

Fahle, M. (2004). Perceptual learning: A case for early selection. *Journal of Vision, 4*(10):4, 879–890, http://www.journalofvision.org/content/4/10/4, doi:10.1167/4.10.4. [PubMed] [Article]

Fiorentini, A., & Nicoletta, B. (1980). Perceptual learning specific for orientation and spatial frequency. *Nature, 287,* 43–44.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research, 51,* 771–781.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature, 402,* 176–178.

Gold, J., Sekuler, A. B., & Bennett, P. J. (2004). Characterizing perceptual learning with external noise. *Cognitive Science, 28,* 167–207.

Green, D. (1964). Consistency of auditory detection judgments. *Psychological Review, 71,* 392–407.

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Karni, A., & Sagi, D. (1993). The time course of learning a visual skill. *Nature, 365,* 250–252.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception,* 36(ECVP Abstract Suppl.).

Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision, 8*(16):10, 1–19, http://www.journalofvision.org/content/8/16/10, doi:10.1167/8.16.10. [PubMed] [Article]

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R*. New York: Springer.

Kurki, I., Hyvärinen, A., & Laurinen, P. (2006). Collinear context (and learning) change the profile of the perceptual filter. *Vision Research, 46,* 2009–2014.

Li, R. W., Klein, S. A., & Levi, D. M. (2008). Prolonged perceptual learning of positional acuity in adult amblyopia: Perceptual template retuning dynamics. *Journal of Neuroscience, 28,* 14223–14229.

Li, R. W., Levi, D. M., & Klein, S. A. (2004). Perceptual learning improves efficiency by retuning the decision "template" for position discrimination. *Nature Neuroscience, 7,* 178–183.

Lu, Z. -L., & Dosher, B. A. (2008). Characterizing observers using external noise and observer models: Assessing internal representations with external noise. *Psychological Review, 115,* 44–82.

Macmillan, N., & Douglas, C. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Mollon, J., & Danilova, M. (1996). Three remarks on perceptual learning. *Spatial Vision, 10,* 51–58.

Murray, R. F. (2011). Classification images: A review. *Journal of Vision, 11*(5):2, 1–25, http://www.journalofvision.org/content/11/5/2, doi:10.1167/11.5.2. [PubMed] [Article]

Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2002). Optimal methods for calculating classification images: Weighted sums. *Journal of Vision, 2*(1):6, 79–104, http://www.journalofvision.org/content/2/1/6, doi:10.1167/2.1.6. [PubMed] [Article]

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

Peterson, M. F., Abbey, C. K., & Eckstein, M. P. (2009). The surprisingly high human efficiency at

learning to recognize faces. *Vision Research, 49,* 301–314.

Peterson, W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Proceedings of the IRE Professional Group on Information Theory, 4,* 171–212.

Petrov, A. A., Dosher, B. A., & Lu, Z. -L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review, 112,* 715–743.

Saarinen, J., & Levi, D. M. (1995). Perceptual learning in vernier acuity: What is learned? *Vision Research, 35,* 519–527.

Sigman, M., & Gilbert, C. D. (2000). Learning to find a shape. *Nature Neuroscience, 3,* 264–269.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

Tjan, B. S., & Legge, G. E. (1998). The viewpoint complexity of an object-recognition task. *Vision Research, 38,* 2335–2350.

Trenti, E. J., Barraza, J. F., & Eckstein, M. P. (2010). Learning motion: Human vs. optimal Bayesian learner. *Vision Research, 50,* 460–472.

Watanabe, T., Náñez, J. E., & Sasaki, Y. (2001). Perceptual learning without perception. *Nature, 413,* 844–848.

Westheimer, G. (1982). The spatial grain of the perifoveal visual field. *Vision Research, 22,* 157–162.

Xiao, L., Zhang, J., Wang, R., & Klein, S. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology, 18,* 1922–1926.

Zhang, Y., Pham, B. T., & Eckstein, M. P. (2006). The effect of nonlinear human visual system components on performance of a channelized Hotelling observer in structured backgrounds. *IEEE Transactions on Medical Imaging, 25,* 1348–1362.

Appendix A

Comparison of GLM and weighted sums for estimating classification images

Simulated data

We ran a computer simulation to compare the efficiency of two classification image estimation methods for one interval rating scale experiment: weighted sums (Murray et al., 2002) and GLM (Knoblauch & Maloney, 2008).

We used a 16-element stimulus with random weights, masked with random noise. Each simulation had the same number of trials (800) as the original experiment. We tested how accuracy is affected by the amount of internal noise (internal-to-external noise ratio γ 1 or 2; close to γ estimated from the double-pass data) and the number of orientation stimulus strength (orientation differences) levels (one or five), even when this was always fixed to five in the experiment. In the simulation, we used the same GLM model and estimation procedure as in the experiment. Estimation error was quantified by comparing the Pearson cross-correlation between the estimated and true templates.

Empirical data

We used both the GLM and the weighted sums method to analyze the empirical data in the experiment. With the weighted sums method, classification images are formed by subclassification images of noise fields averages of each response $\times$ target stimulus combination. We used the optimal weighting that maximizes the expected SNR, taking into account both rating scale response distribution and target detectability at each orientation target level (Murray et al., 2002). For those observers who did not consistently use all eight rating possibilities, neighboring responses were pooled together to reduce effective response ratings. This step was used for the data for both methods.

Results

Figure A1 shows the results of the simulation, quantified as mean correlation error between the true and estimated templates. The left side of the figure is from a simulation where internal-to-external noise ratio $\gamma = 1$ and right side \ $\gamma = 2$. GLM outperforms weighted sums in both conditions; the difference is most pronounced with less internal noise. Figure A2 shows the empirical data; GLM is plotted on the right and weighted sums classification image is plotted on the left. Shapes of classification images are similar but GLM results have less interindividual variance, suggesting that they contain less estimation error.

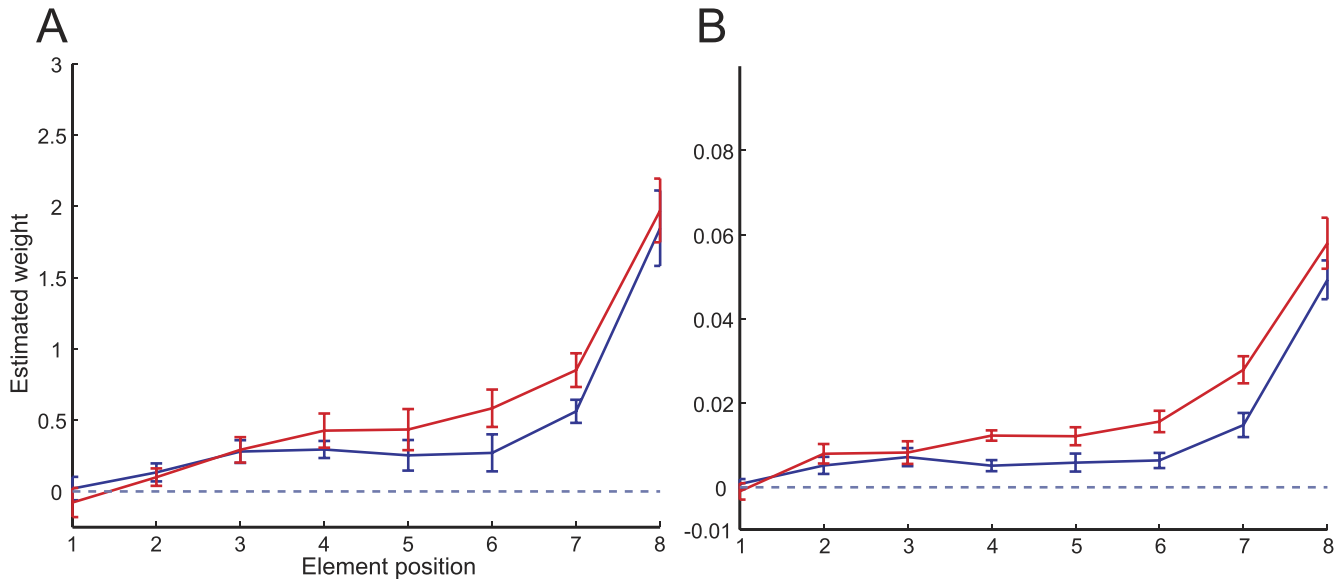

Figure A1. GLM versus weighted sums simulation results. Bars represent mean correlation error between true template and estimated template. Purple bars: weighted sums, green bars: GLM. Bar groups represent different number of criteria. Upper row: Simulation with 1 MOCS signal level. Lower row: Simulation with five levels (800 trials; 160 trials/level). Error bars represent 1 *SEM*.

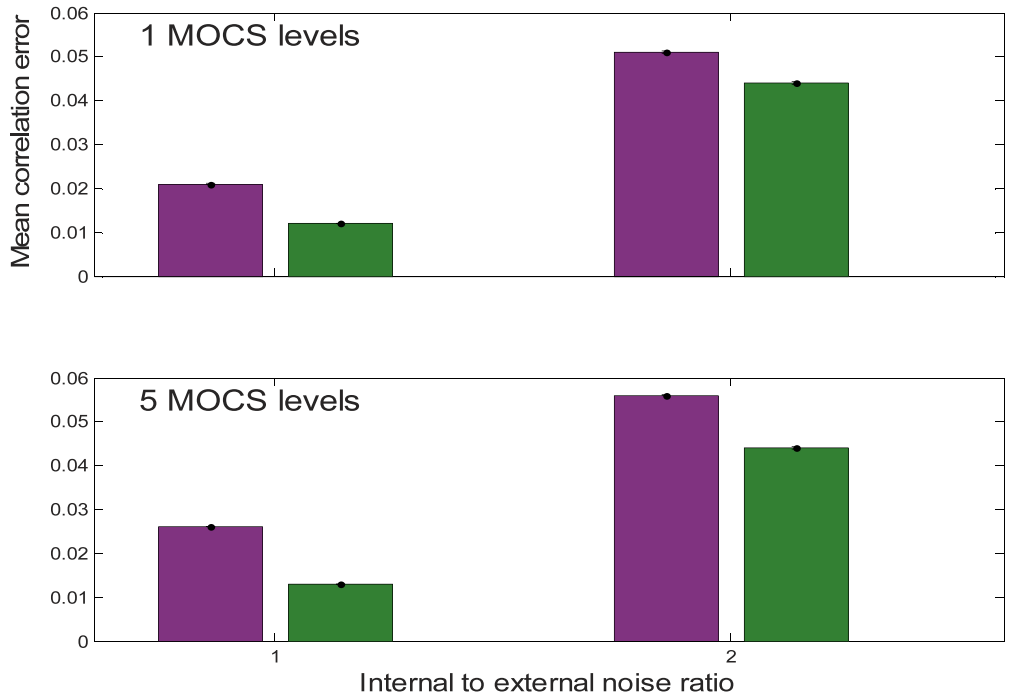

Figure A2. Comparison of GLM and weighted sums classification image methods with empirical data. (A) Averages of classification images estimated using weighted sums (Murray et al., 2002). (B) Averages of classification images using GLM (Knoblauch & Maloney, 2008). The red curve is the average classification image for the first session; the blue curve is the average classification image for the last session. Error bars represent 1 *SEM*. Shapes of classification images are similar but GLM results have less interindividual variance, suggesting that they contain less estimation error.

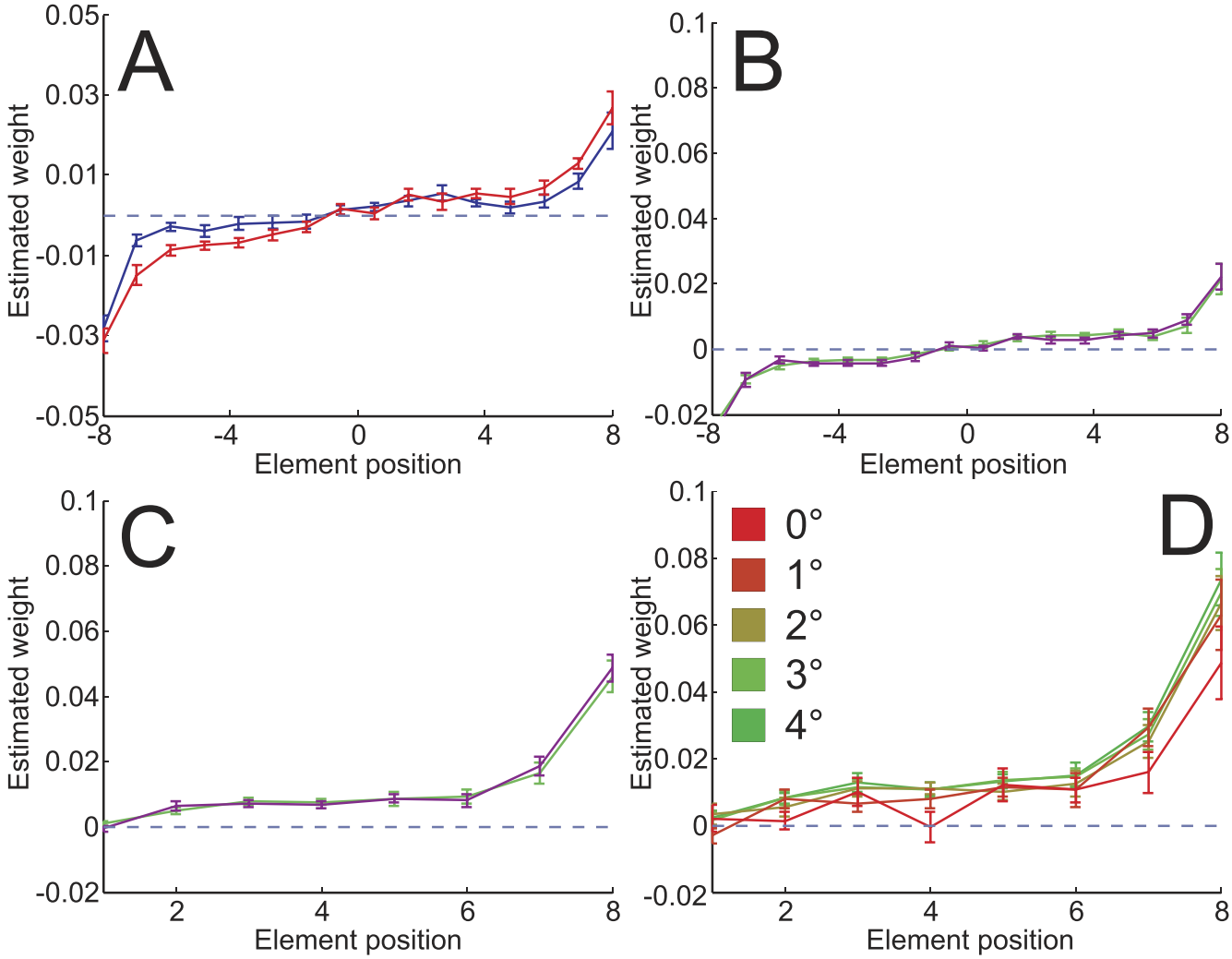

Figure A3. Control analyses. (A) Classification images without spatial averaging across the top and mirror-bottom of the stimulus. Average classification image (10 subjects) for the first session (blue) and the last session (red). Estimated template weights are plotted against element position (x-axis). (B) Classification images without spatial averaging were analyzed separately for trials where the target was tilted toward the center of the screen (green curve) and trials where the target was tilted away from the center (blue curve). (C) Classification images were analyzed separately for the left-tilted target (green curve) and the right-tilted target (blue curve). All subjects, levels, and trials were pooled together. (D) Classification images were analyzed separately for each of five target "baseline" tilt levels; red = no tilt; green = maximum tilt. Classification images are an average of all 10 observers and all five sessions.

Appendix B

We ran several control analyses in order to justify averaging the mirror-symmetric parts of classification images as well as averaging classification images for left/right responses. Moreover, we investigated whether the baseline tilt had an effect on classification images. First, we analyzed classification images without averaging the mirror symmetric parts. The average classification image (over 10 subjects) and individual data (not shown) show that learning effects were similar in both parts of the template (Figure A3A). We compared the difference between the top and mirror-symmetric bottom classification images (across all sessions to improve the statistical power), but we did not find any significant difference, Hotelling $T^2(8, 9) = 14.29$, $p = 0.845$. Next, as six observers had the stimulus on the left side and four had the stimulus on the right side, we compared whether average classification images were different in trials where the target was tilted toward the center of the screen or away from it (Figure A3B). In addition, we asked whether there was any difference in processing the left and right tilt (Figure A3C). Differences were minute and statistically nonsignificant: toward center/away difference $T^2(8, 9) = 384.45$, $p = 0.09$; left/right difference $T^2(8, 9) = 65.54$, $p = 0.4$. We then compared classification images at different levels of baseline target tilt from the average classification images (collapsed across all sessions and observers to

increase statistical power). Overall, classification images are similar in both shape and amplitude. However, the amplitude of the maximum (4°) tilt level was a bit lower (see Figure A3D). The difference was not significant when compared with the minimum tilt level, $T^2(8, 9) = 291$, $p = 0.12$. We then compared classification images between all tilt level combinations, but no significant differences were found (when corrected for multiple comparisons using false discovery rate statistics at $\alpha = 0.05$).