Title
Augmented Hebbian reweighting accounts for accuracy and induced bias in perceptual learning with reverse feedback

Journal
Journal of Vision, 15(10)

ISSN
1534-7362

Authors
Liu, Jiajuan
Dosher, Barbara Anne
Lu, Zhong-Lin

Publication Date
2015-09-29

DOI
10.1167/15.10.10

Peer reviewed

Augmented Hebbian reweighting accounts for accuracy and induced bias in perceptual learning with reverse feedback

Jiajuan Liu
Department of Cognitive Sciences, University of California, Irvine, CA, USA

Barbara Anne Dosher
Department of Cognitive Sciences, University of California, Irvine, CA, USA

Zhong-Lin Lu
Department of Psychology, The Ohio State University, Columbus, OH, USA

Using an asymmetrical set of vernier stimuli (−15″, −10″, −5″, +10″, +15″) together with reverse feedback on the small subthreshold offset stimulus (−5″) induces response bias in performance (Aberg & Herzog, 2012; Herzog, Eward, Hermens, & Fahle, 2006; Herzog & Fahle, 1999). These conditions are of interest for testing models of perceptual learning because the world does not always present balanced stimulus frequencies or accurate feedback. Here we provide a comprehensive model for the complex set of asymmetric training results using the augmented Hebbian reweighting model (Liu, Dosher, & Lu, 2014; Petrov, Dosher, & Lu, 2005, 2006) and the multilocation integrated reweighting theory (Dosher, Jeter, Liu, & Lu, 2013). The augmented Hebbian learning algorithm incorporates trial-by-trial feedback, when present, as another input to the decision unit and uses the observer's internal response to update the weights otherwise; block feedback alters the weights on bias correction (Liu et al., 2014). Asymmetric training with reversed feedback incorporates biases into the weights between representation and decision. The model correctly predicts the basic induction effect, its dependence on trial-by-trial feedback, and the specificity of bias to stimulus orientation and spatial location, extending the range of augmented Hebbian reweighting accounts of perceptual learning.

Introduction

In perceptual learning, performance improves with practice by improving the sensitivity to or discrimination between stimuli (Fahle & Poggio, 2002; Lu & Dosher, 2012; Sagi, 2011; Sasaki, Nanez, & Watanabe, 2010). However, the training experiences during perceptual learning in some circumstances may also alter or shift the apparent response biases in addition to improving sensitivity (Jones, Moore, Amitay, & Shub, 2013; Wenger & Rasche, 2006). Although most experimental training protocols are designed with balanced stimulus presentations and accurate feedback, other training experiences are designed to induce systematic tendencies toward biased responses. Furthermore, relative stimulus frequencies and accurate feedback may not occur in real-world environments. A series of recent studies by Herzog and colleagues (Aberg & Herzog, 2012; Herzog, Eward, Hermens, & Fahle, 2006; Herzog & Fahle, 1999) explored the consequences for sensitivity and bias in perceptual learning using asymmetric training sets and reverse (false) feedback under varied feedback conditions. Their basic induction protocol trained line offset vernier stimuli and used trial-by-trial feedback on errors. Reverse feedback on a single subthreshold offset stimulus (i.e., −5″ left) in the presence of balanced superthreshold left and right offset stimuli was sufficient to introduce systematic biases in responding, shifting responses to all stimuli in the direction of the reversed (false) feedback. (In this notation, the ″ stands for arcseconds of visual angle, and the value is negative where the bottom line is to the left of the top line.) Understanding the mechanisms of perceptual learning in these imbalanced training environments poses a challenge to theories and models of perceptual learning. Accounting for these results may be an important step in generalizing training protocols to practical applications where the training experiences may incorporate imbalanced stimulus frequencies and occasional misleading feedback.

Citation: Liu, J., Dosher, B. A., & Lu, Z.-L. (2015). Augmented Hebbian reweighting accounts for accuracy and induced bias in perceptual learning with reverse feedback. *Journal of Vision, 15*(10):10, 1–21, doi:10.1167/15.10.10.

doi: 10.1167/15.10.10 Received February 28, 2015; published September 29, 2015 ISSN 1534-7362

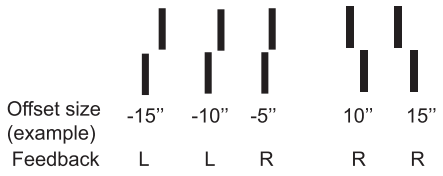

Figure 1. The basic asymmetric vernier training stimulus set with reversed (false) feedback indicated. Only one vernier stimulus is shown on any trial.

Some of the classic models of perceptual learning (Herzog & Fahle, 1998; Poggio, Fahle, & Edelman, 1992; Weiss, Edelman, & Fahle, 1993) were developed to account for perceptual learning in these hyperacuity tasks. These articles showed that a variety of learning algorithms could account for improvements in performance with practice, including in some cases unsupervised learning (Weiss et al., 1993), and identified challenges in modeling perceptual learning (Herzog & Fahle, 1998). The models have not been systematically applied to the induced bias problem.

In this article, we examine the ability of the fully implemented augmented Hebbian reweighting model (AHRM) of perceptual learning and its extensions to account for the data in the induced bias paradigms of Herzog and colleagues. The AHRM is a computationally implemented perceptual learning model that takes stimulus images as inputs, generates choice responses on each trial, and uses a Hebbian learning rule that is augmented by feedback and bias control to improve performance. The AHRM model was initially developed to account for perceptual learning in nonstationary environments with biased external noise (Petrov, Dosher, & Lu, 2005, 2006). The AHRM has also been used to model the mechanisms of perceptual learning (Lu, Liu, & Dosher, 2010), the effectiveness of training in different difficulty levels (Liu, Lu, & Dosher, 2010, 2012), and feedback effects (Liu, Dosher, & Lu, 2014; Petrov et al., 2005, 2006). Most recently, Liu et al. (2014) extended the AHRM to account for learning curves under different forms of feedback, including no-feedback, false-feedback, block-feedback, and trial-by-trial feedback experiments (Herzog & Fahle, 1997; for a review see Dosher & Lu, 2009). Dosher, Jeter, Liu, and Lu (2013) developed an integrated reweighting theory (IRT) that extended the AHRM by using multilevel representations to account for specificity and transfer over retinal locations.

The data on asymmetric training and reversed feedback provide new tests of these models and the hypothesis that much of perceptual learning is achieved through reweighting—the changed readout of evidence from stable perceptual representations (Dosher & Lu, 1998, 1999). We show that the reweighting theories provide an excellent account of perceptual learning in asymmetric and biased training protocols, accounting for many of the phenomena in these interesting studies and extending the application range of the theoretical framework of the AHRM and IRT.

Induction of response bias by reverse feedback in asymmetric sets

Figure 1 illustrates the bias-induction training protocol from the studies of perceptual learning in vernier judgments (i.e., Herzog & Fahle, 1999), which includes asymmetric training sets and reverse feedback. In the basic design, the vernier training set consists of stimuli with two bars in which the bottom bar is aligned slightly to the left or right of the top bar—here (in different trials) by $-15''$, $-10''$, $-5''$, $+10''$, or $+15''$. Feedback is accurate for the medium and large (suprathreshold) offsets, but feedback is reversed for the smallest (subthreshold) offset, which is objectively shifted left by a very small amount while the feedback specifies a *right* response. Variations on the basic training set used different relative proportions of the stimuli during training and different proportions of feedback reversal (Herzog & Fahle, 1999) and showed sensitivity to the specific asymmetric and reversed feedback training.

Training distinct and separate bias patterns for stimuli and judgments using different orientations or at different spatial positions demonstrated a remarkable specificity of these effects (Herzog et al., 2006). Different types and sequences of feedback were differentially effective in training bias, leading to the conclusion that training sensitivity and training biases show different and interacting results (Aberg & Herzog, 2012). Finally, the prior literature makes some claims about differential consolidation of sensitivity and bias criterion effects (Aberg & Herzog, 2012).

Augmented Hebbian reweighting and perceptual learning

Dosher and Lu (1998, 1999) proposed that many perceptual learning phenomena could be modeled through reweighting sensory evidence to a decision (Dosher & Lu, 1998, 1999). This principle has been implemented within the context of an AHRM (Petrov et al., 2005, 2006; see also Dosher et al., 2013; Liu et al., 2010, 2012, 2014). Perceptual learning occurs through reweighting of evidence from stable perceptual representations to a decision structure for a specific perceptual task (Dosher & Lu, 2009). The reweighting model for perceptual learning consists of a representation module, or visual front end, and a decision module that takes the activities of representation units and generates a response. A learning module alters the weights of the connection between the representation system and decision using Hebbian learning augmented by feedback, when it is present, and a criterion-control

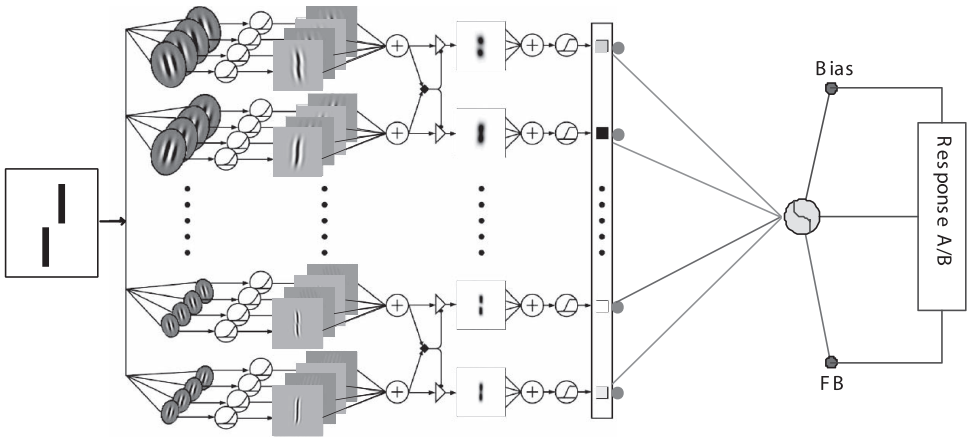

Figure 2. An overview of the AHRM. There are three main subsystems or modules of the AHRM framework: a representation subsystem, a decision subsystem, and a learning subsystem. In this application, the representation subsystem computes the activations of spatial frequency and orientation tuned filters, subject to normalization gain control and internal noise; the decision subsystem takes the weighted sum of activation outputs from the representation subsystem, adds in the bias term and the decision noise, and uses a nonlinear decision unit to classify the stimulus and generate a response; and the learning subsystem incorporates the feedback into a revised decision variable and updates the weights between the representation units and decision through Hebbian learning. The model takes individual stimuli as inputs, generates responses, and learns in an exact simulation of the experimental protocol experienced by the observer in a perceptual learning experiment.

unit that tracks the balance of responses in an experimental situation. Figure 2 illustrates the structure of the AHRM. Accounting for performance and learning in any particular task requires a representation system that codes the information relevant to that task. For example, motion tasks require a front-end representation module that extracts motion from the stimulus (Lu et al., 2010; Tlapale, Dosher, & Lu, 2013, 2014). A Hebbian learning rule incrementally updates the weights between the representation system and the decision system, although other learning systems have also been proposed (i.e., Jacobs, 2009). The model exactly replicates the sequences of training trials experienced by the observer to generate behavioral predictions for each training protocol.

The studies by Herzog and colleagues considered in the current simulation study used line offset vernier stimuli. As in our previous simulation (Liu et al., 2014) of the effects of different feedback manipulations on learning line offset vernier judgments (Herzog & Fahle, 1997), we used a representation module with a set of spatial frequency and orientation tuned units, originally used to code oriented Gabor stimuli (Petrov et al., 2005, 2006) and used for modeling tilt judgments of noisy line grids (Jacobs, 2009). This choice reflects the similarity between line offset vernier and orientation judgments (Saarinen & Levi, 1995; Weiss et al., 1993). (Elsewhere we used radial basis location coding as the representation to model perceptual learning in three-dot vernier offset and bisection studies; Huang, Lu, & Dosher, 2012.) The IRT (Dosher et al., 2013) extended the reweighting framework by incorporating multilevel representations to account for transfer over retinal positions.

The orientation–spatial frequency representation module and the augmented Hebbian learning module are based on those in previous articles (Liu et al., 2010, 2012, 2014; Lu et al., 2010; Petrov et al., 2005, 2006). A short description of the model equations is provided in Appendix A for convenience. The representational subsystem consists of units or channels that are spatial frequency and orientation selective, centered on five spatial frequencies and seven (or 12) orientations that span the orientation space. Each unit incorporates nonlinearity, normalization, and stochastic internal noise in the response and includes the response to external noise in paradigms that incorporate external noise in the stimulus (Dosher et al., 2013; Liu et al., 2010, 2012; Lu et al., 2010; Petrov et al., 2005, 2006). Many of the parameters, such as orientation and spatial frequency bandwidths of the representation units, are set a priori based on the physiology and prior applications; the nonlinearity and normalization are broadly consistent with normalization and gain control observed in neural systems (Carandini, Heeger, & Movshon, 1997; Heeger, 1992). This front end is designed to incorporate many known properties of early visual processing, which is important for predicting performance for stimuli of different contrasts or different external noise. The vernier stimuli in the studies by Herzog and colleagues do not incorporate either contrast variation or external noise. We use the full model for consistency and to allow generalization to other kinds of experiments.

The induced biases and improvements in sensitivity that result from asymmetric training and reversed feedback under different feedback protocols provide a

new challenge and test for the reweighting framework and the AHRM model.

Method

AHRM model and simulation methods

The AHRM was implemented in MATLAB. On each trial, the model processes grayscale images of the experimental stimuli as inputs and generates binary (*left–right*) responses as outputs. It learns on each trial by adjusting the connection weights between the representation units and the decision unit. The model replays each experimental protocol, including the number of trials of each kind of stimulus, the nature of feedback, and the number of training sessions—that is, it reprises the experimental protocols experienced by the observers. Each simulated experiment was repeated 1,000 times to generate the predictions of the model.

Most of the parameters of the model are set a priori. In particular, most front-end or representational subsystem parameters were set from the physiological literature or fixed based on model fits to experimental data in a number of other applications (Dosher et al., 2013; Liu et al., 2010, 2012; Lu et al., 2010; Petrov et al., 2005, 2006; see Appendix A). Prior knowledge about orientation was embodied in initial weights (priors) set proportional to the preferred orientation of the units $w_i = (\theta_i / 45)w_{init}$. These parameters were held constant over all of the simulations reported here.

A small number of variable parameters were chosen to approximate the overall performance level and learning rates of the target data set. These include the internal multiplicative noise σ_m, decision noise σ_d, scaling factor a, the weight on feedback w_f, and (model) learning rate η. The learning rate and weights on the bias and feedback units were adjusted to approximate the pattern of learning and bias in the data. Detailed optimization of the fits of the model to the data, carried out in some of our previous articles (Liu et al., 2010, 2012; Lu et al., 2010), are extremely time consuming—sometimes taking months of grid search computations to yield just slightly better fits. However, many regions of the parameter space generate predictions consistent with the qualitative properties of the observed data pattern(s). Here we perform only approximate fits of model to data in order to enable us to examine a wider range of experimental findings. The match between model and data is indexed by the rank-order correlation between the model predictions and the observed data points, Kendall's τ (Kendall, 1938; Kendall & Gibbons, 1990)—a measure of concordance between data and model predictions that is relatively robust to distributional issues (Newson, 2002):

$$\tau = \frac{(n_C) - (n_D)}{1/2(n)(n-1)},$$

where n_c is the number of data pairs with concordant order between the two data sets and n_D is the number of data pairs with discordant order between the two data sets. We also report the parametric estimate of the proportion of variance accounted for by the model

$$r^2 = 1 - \frac{\sum_{i=1:n}(x_i - \hat{x}_i)^2}{\sum_{i=1:n}(x_i - \bar{x})^2},$$

where x_i is an observed data value, $\hat{x}_i$ is the corresponding value predicted by the model, and $\bar{x}$ is the mean of the observed data. Since we do not carry out precise quantitative fits of the model to data, the values of the proportion of variance accounted for by the model are lower than if we had. For a number of experiments, these values were also limited by the small range and noise in the behavioral data sets.

Experimental data sets

In this article, we selected representative experimental data from the three induced bias articles of Herzog and colleagues (Aberg & Herzog, 2012; Herzog et al., 2006; Herzog & Fahle, 1999). These experiments began with an initial assessment of threshold for each observer. The largest offset condition was set to be suprathreshold, while the medium and small offsets were slightly below and clearly below threshold; the large offsets were generally about three times the magnitude of small offsets. The values (i.e., −15″, −10″, −5″, 10″, and 15″) are listed as examples (see Figure 1). In many of these experiments, the single reverse-feedback condition was presented with probability 1/3. Herzog and Fahle (1999) included five vernier offsets in their experiments. Herzog et al. (2006) simplified the experiment to include only three vernier offsets (i.e., −15″, −5″, and +15″). Feedback is presented on errors for trial-by-trial conditions, while overall accuracy is provided at block breaks for block-feedback conditions. More details about each modeled experiment are provided as they are treated.

Results

Inducing bias with asymmetric training and false feedback

Herzog and Fahle (1999) were the first to introduce a perceptual learning protocol specifically designed to

induce biases as well as changes in discrimination sensitivity. Their paradigm presented the asymmetric stimulus set of line offset vernier stimuli that shifted the bottom line left or right by medium or large shifts but added (false) reverse feedback to a singleton small left stimulus (or vice versa), where each stimulus offset is tested with some frequency. The net effect is to increasingly shift toward *right* responses—as though the misleading feedback on this relatively ambiguous (below-threshold offset) stimulus induces observers to lower their criterion for a *right* response.

We simulated the data of experiment 3 of Herzog and Fahle (1999). Other experimental variants in their article manipulated the relative frequencies of the five stimuli (i.e., $-15''$, $-10''$, $-5''$, $+10''$, and $+15''$) and the frequency of reversed feedback for the unique small offset stimulus ($-5''$) in various ways. We chose experiment 3 because it used a design with 1/3 probability of the reversed-feedback condition, which is typical of the designs in subsequent articles, and because the data were representative of the magnitude of the basic effect across multiple experiments.

This training affected performance on the small offset stimuli that received reversed feedback but also affected performance for the medium ($\pm 10''$) and large ($\pm 15''$) offset stimuli. Figure 3 shows the stimuli (Figure 3a), the biased response data from the experiment of Herzog and Fahle (1999; Figure 3c), and corresponding predictions of the AHRM model (Figure 3d). It also shows the evidence (activity) in the orientation and spatial frequency tuned units (filters) for these stimuli (Figure 3b).

As seen by comparing the model predictions in Figure 3d with the data in Figure 3c, the AHRM model naturally accounts for the biased responding induced by these asymmetric training paradigms. Reversed feedback on the $-5''$ stimulus results in the observer learning to shift toward feedback-consistent *right* responses. This increasingly reduces the percentage correct for the left offset stimuli (left offset stimuli were the only data shown in the original article). Then, when the reversed feedback is replaced with accurate feedback (at the vertical dashed line), performance shifts rapidly toward more *left* responses and therefore increased percentage correct on left offset stimuli. The model predictions qualitatively replicate the pattern in the data; they are rank-order consistent with the observed data (Kendall's $\tau = 0.692$, $p \ll 0.001$). The proportion variance accounted for by the model is $r^2 = 0.636$ ($p < 0.01$). Model parameters, selected to approximately mimic the levels in observed data, are listed in Table 1. A time-intensive grid search on model parameters and the corresponding simulated results would almost surely provide a slightly improved detailed fit to the data, but the predicted ordinal pattern occurs over most of the model parameter space.

The multiple experiments in the original article (Herzog & Fahle, 1999) showed a similar general pattern: With different probabilities of the smallest left offset with reversed feedback (labeled *right*), performance of all left verniers dropped and then quickly rebounded with the introduction of correct feedback. Also, if the probability of reversed feedback is higher, the biasing effect is more prominent. We show the AHRM prediction for experiment 3, but the AHRM is broadly consistent with all the experiments in the study.

The AHRM model accounts for these results in the following way. During bias induction training, reverse feedback for the small offset-left stimulus shifts the weights toward the *right* response. The reverse feedback for this small offset drives the postsynaptic activity at the decision unit toward the incorrect response, shifting weights to favor the rewarded response through Hebbian learning. These changes are concentrated in the weights for orientation channels near the vertical (0°, +15°, and −15°) that are most sensitive to the very small angles of the vernier stimuli. Subsequent training with accurate feedback shifts weights to favor the now-dominant *left* feedback.

In the AHRM account, then, the induced biases seen in the empirical data are predominantly encoded by reweighting the activity in the sensory representations into the decision unit toward the dominant feedback category—a learned weight-encoded bias. In fact, the bias control unit of the AHRM, which seeks to balance the two responses in the recency-weighted response history, tends to moderate or oppose induced weight changes that lead to an imbalanced response history. Examples of changes in weight structures of the AHRM are displayed and discussed in more detail in Appendix B.

Induced biases depend on trial-by-trial feedback

Aberg and Herzog (2012) examined the extent to which the perceptual learning–induced biases in discrimination, first documented in Herzog and Fahle (1999), reflect the trial-by-trial feedback of those studies. They compared the induced bias effect in separate groups of observers trained in the trial-by-trial reversed feedback protocol of their original demonstrations, a trial-by-trial correct-feedback condition (that eliminated the reversed feedback for the $-5''$ offset stimuli), seven-trial blocked correct feedback, seven-trial blocked reversed feedback, 84-trial blocked reversed feedback, and no feedback. In the blocked conditions, the corresponding aggregate accuracy information was presented at the end of blocks of the specified length. Each of the six feedback groups was first trained with the assigned feedback, followed by three blocks without feedback, and then three blocks

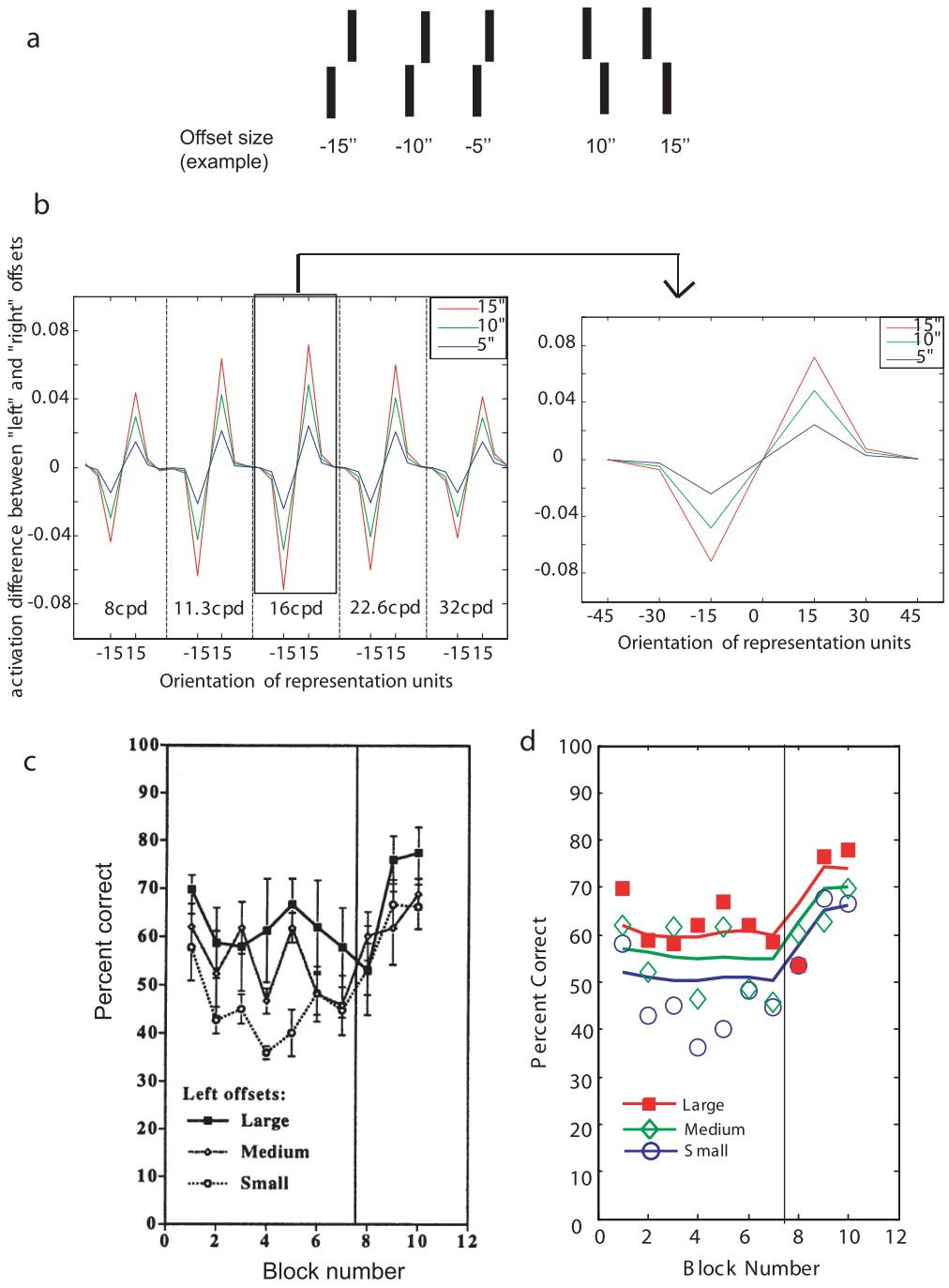

Figure 3. Perceptual learning of biased responding through practice on an asymmetric stimulus set and selective reverse feedback in a line offset vernier task and predictions of the AHRM reweighting model of perceptual learning. (a) The asymmetric stimulus set used in the experiment. (b) The differential evidence in representation units of the AHRM representation system for these stimuli. (c) Percentage correct for left offset stimuli from Herzog and Fahle (1999, experiment 3, figure 9; corresponding data for right offset stimuli not shown). The vertical line corresponds to an elimination of reverse feedback for the −5″ stimulus. (d) Corresponding pattern of bias predictions from the AHRM model (parameter values from Table 1; see text) in lines and data in symbols. Panel c is adapted with permission from Herzog, M.H., & Fahle, M. (1999). “Effects of biased feedback on learning and deciding in a vernier discrimination task.” *Vision Research*, *39*(25), 4232–4243.

with accurate feedback. The trial-by-trial reversed-feedback condition is a near replication of the pattern shown in Herzog and Fahle (1999), with several (three) no-feedback training blocks interposed between the induction training and the final correct feedback blocks.

Figure 4 shows the original data display, an alternative graphing of the data that better illustrates

Variable	Parameter	Value			
Parameters set a priori	Orientation spacing	$\Delta\theta = 15^\circ$			
	Spatial frequency spacing	$\Delta f = 0.5$ octave			
	Maximum activation level	$A_{max} = 1$			
	Weight bounds	$w_{min/max} = \pm 1$			
	Running average rate	$\rho = 0.02$			
	Activation function gain	$\gamma = 0.8$			
	Bias weight	$w_b = (2*pc - 1)*w_{bf}$			
	Normalization constant	$k = 0$			
	Internal additive noise	$\sigma_1 = 0$			
	Initial weight scaling factor	$w_{ini} = 0.169$			
Parameters constrained by published data	Orientation tuning bandwidth	$h_\theta = 30^\circ$			
	Frequency tuning bandwidth	$h_f = 1.0$ octave			
	Radial kernel width	$h_r = 2.0^\circ$ of visual angle			
Parameters optimized to fit the present data	Feedback weight	$w_f = 0.2$			
	Bias control weight factor	$w_{bf} = 0.04$			
		Study 1	Study 2	Study 3	Study 4
	Representation scaling factor	$a = 0.4$	0.4∼0.45	0.3	0.3
	Internal multiplicative noise	$\sigma_m = 0.01$	0.01∼0.03	0.02	0.02
	Decision noise	$\sigma_d = 0.024$	0.012∼0.016	0.015	0.01
	Learning rate	$\eta =$ 5e-4	3e-4	4e-4	3e-4

Table 1. Model parameters.

the relation to stimulus variation, and predictions of the AHRM model. Figure 4a duplicates the data graph from Aberg and Herzog (2012, figure 2), which graphed performance as the hit rate (left stimuli categorized as *left* and right stimuli categorized as *right*) for each stimulus type (labeled *big left*, *middle left*, *small left*, *middle right*, and *big right*). Figure 4b shows the same data as percentage *right* in order to more clearly show the separation of performance for the five vernier offsets and more clearly track the overall shifts toward *left* or *right* responses both in the data and in the model predictions. Figure 4e duplicates the data from Aberg and Herzog (2012, figure 3) for derived criteria from signal detection analysis, and Figure 4f shows the AHRM predictions.

The AHRM model generates predictions that generally parallel the effects of perceptual training under the six feedback conditions. It replicates the original learning effects with induced biases by trial-by-trial feedback with reversed feedback for the small left condition (top right subpanel). This shifts the performance upward, toward *right* responses, corresponding with the distribution of feedback. The AHRM correctly predicts shifts downward (toward increased *left* responses) for the trial-by-trial correct-feedback condition (upper middle subpanel), in which *left* feedback dominates, reflecting the relative frequencies of the stimuli. The model also correctly predicts weak or nonexistent induced bias in the no-feedback and various blocked-feedback conditions.

Figure 4d shows the relationship between the observed data and the model predictions for each training feedback group in separate subpanels. Note that the variation in observed performance (Figure 4d) for each stimulus over the course of learning partly reflects binomial variability in observed proportions given the sample sizes in the experiment. The model predicts upward shifts (more *right* responses) for the reverse-feedback conditions and downward shifts (more *left* responses) for the correct-feedback conditions. Other than minor adjustments in the initial scaling constants (*a*s) and noise terms used to approximately match performance in the initial training blocks, the same model parameters were used to predict performance in the different feedback groups. Figure 4e and f shows criterion estimates derived from a signal detection analysis for the data and the model; the data show smaller deviations between the derived criteria of the reversed and accurate trial-by-trial feedback conditions at a single point in block 11, not mirrored in the model predictions (for discussion, see Aberg & Herzog, 2012). This issue is not seen in corresponding fits to the percentage *right*. (The model predictions are a better match to criterion estimates from all data rather than from the subthreshold singleton used in Figure 4e by Aberg and Herzog [2012].)

The AHRM simulation accounts for the data pattern qualitatively; the approximate fit also does a relatively good job quantitatively. The simulated model predictions are rank-order consistent with the observed data (Kendall's $\tau = 0.772$, $p \ll 0.01$). The proportion

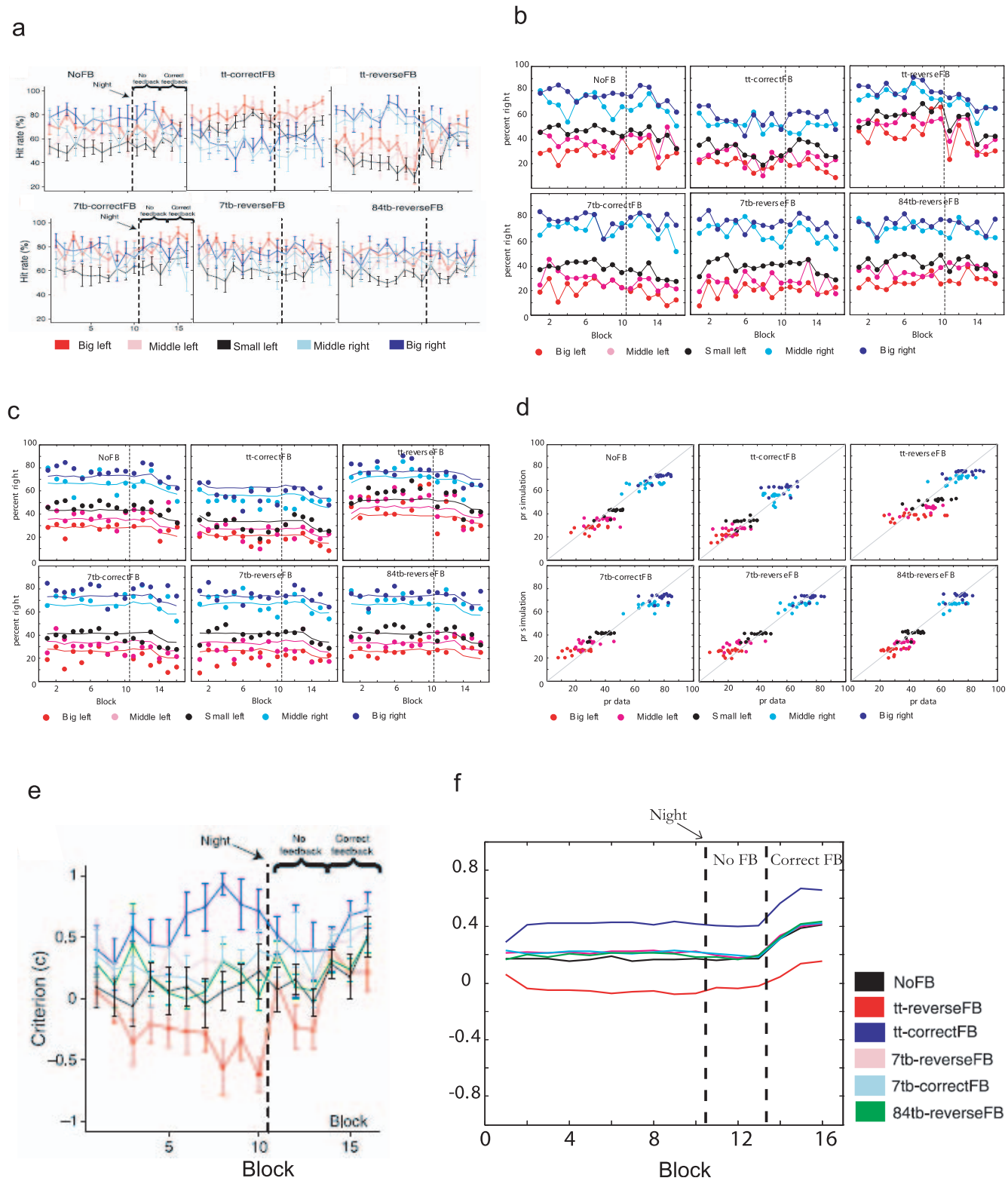

Figure 4. The dependence of induced bias in asymmetric training protocols on the type of feedback used in training and predictions of the AHRM. The stimulus sets are the same as those in Figure 3. (a) Performance (percentage hits) in six feedback conditions (no feedback, trial-by-trial correct feedback, trial-by-trial reversed feedback, seven-trial blocked correct feedback, seven-trial blocked reversed feedback, and 84-trial blocked reversed feedback) from Aberg and Herzog (2012, figure 2). (b) The data in panel a graphed as percentage *right* reveals the different response rates for big left, middle left, small left, middle right, and big right stimuli. (c) The predictions of the AHRM model for the six feedback groups (lines) along with the experimental data (dots). (d) The relationship between the observed percentage *right* and the AHRM predictions. (e) Derived measures of average decision criteria as a function of block of training for the six feedback groups from Aberg and Herzog (2012, figure 3). (f) Predictions of the AHRM model corresponding with the data in panel e. See the text for a discussion. Panels a and e are adapted with permission from Aberg, K.C., & Herzog, M.H. (2012). "Different types of feedback change decision criterion and sensitivity differently in perceptual learning." *Journal of Vision*, *12*(3):3, 1–11.

variance accounted for by the model is $r^2 = 0.883$ ($p \ll 0.01$). Although parameters in the model are varied to get the scale and overall rates, the differential feedback conditions all use the same parameters, with the exception of small differences in scaling factor and noise terms to slightly adjust for minor overall group differences. Model parameters are listed in Table 1. Further search of the parameter space might slightly improve the fit. These model results are consistent with an earlier AHRM account of differential learning when using different forms of feedback in standard (unbiased) vernier perceptual learning (see Liu et al., 2014).

As in the previous analysis of the bias-induction paradigm, the shifts in bias in the AHRM account are encoded into the structure of weights into the decision unit. The false feedback in the trial-by-trial reverse-feedback condition shifts behavior toward a *right* response, while accurate trial-by-trial feedback shifts toward *left* responses. As found previously (Herzog & Fahle, 1997), no-feedback conditions do little to improve performance accuracy. Although block feedback can in some circumstances promote learning (Herzog & Fahle, 1997), in this case it is ineffective.

The model simulations of the trial-by-trial reverse-feedback condition follow those for the previous experiment, as the weights shift in the same direction as the preponderance of the feedback, or *right*. The weights tend not to change substantially in the absence of feedback because the (early) postsynaptic activation at the decision unit tends to be very small for vernier stimuli: The orientations are all very close to vertical or zero. Trial-by-trial accurate feedback shifts weights and performance in favor of the dominant *left* response. Although block feedback can lead to learning in balanced designs, it is not sufficient to induce the overall shifts to *right* or *left* response because only trial-by-trial feedback can shift the postsynaptic weight toward a particular response. See Appendix B for a depiction of the changes in the weight structures and a more detailed discussion.

Overall, then, the AHRM provides an integrated predictive account of the complex pattern of results for different forms of feedback in these asymmetric training conditions with no feedback, reversed and correct trial-by-trial feedback, and various forms of blocked feedback. Although parameters are selected to approximately match the level and scale of the empirical data, the differences between feedback conditions are qualitative predictions of the model. Other interesting issues discussed in Aberg and Herzog (2012), such as those related to overnight consolidation effects and so forth, are not incorporated in the current AHRM; they are considered briefly in the Discussion.

Independent (opposite) induced bias in different trained orientations

The induced biases created by the basic induction paradigm can be specific to the trained stimuli. One interesting demonstration of the specificity of induced bias was experiment 3 in Herzog et al. (2006). This experiment trained either horizontal or vertical offset stimuli with different stimulus–feedback regimes in succession. The theoretical purpose was to examine whether the induced biases are learned separately for the horizontal and vertical stimuli or whether they instead reflect the consequences of perceptual learning on a single shared criterion function for the response. Horizontal and vertical offset stimuli were trained successively in different phases (blocks). The sequence of training phases was designed to induce biases toward opposite response keys in the horizontal and vertical offset stimuli. Initial biases in the other orientation were measured in sessions of balanced testing of very small offset stimuli without feedback. Compared with Herzog and Fahle (1999), Herzog et al. (2006) simplified the stimulus set, which consisted of only a balanced set of large offsets together with a singleton small offset.

Phase H1 tested small ($\pm 5''$) horizontal offsets to measure baseline bias. Phase V1 trained vertical offsets (i.e., $-15''$, $-5''$, and $+15''$) with reverse feedback for the small ($-5''$) offset to induce bias; the small offset was chosen in the direction of "the natural response bias" for each observer (so the bias is somewhat exaggerated). Phase V2 tested the persistence of induced bias without feedback with the same stimulus set (i.e., $-15''$, $-5''$, and $+15''$). Phase V1 was repeated to refresh the induced vertical bias. Phase H1* tested very small ($\pm 1''$) horizontal offsets without feedback—essentially a pure assessment of bias—and found hit rates for both stimuli near 50%. Phase H2 trained horizontal offsets (i.e., $-15''$, $+5''$, and $+15''$) with reverse feedback for the small offset such that the induced bias would shift responses to the opposite response key, as for V1. Phase V3 tested very small ($\pm 1''$) vertical offsets without feedback to document the persistence of (opposite response key) vertical bias. Finally, phase V1 was repeated to show the persistence of the induced vertical bias. In short, this study showed that biases for horizontal or vertical offset stimuli could be induced to favor the opposite keys with little interaction between them. Additionally, there was a tendency for the induced biases to reduce somewhat during no-feedback phases, suggesting a return to balanced responding in the absence of continued false feedback.

Our simulation recapitulated the same series of training blocks. The simulation placed the small offset singleton receiving false feedback in opposite vertical

and horizontal directions (i.e., one was negative and the other was positive).

We simplified the implementation of the simulation by approximating the smallest $\pm 1''$ vernier used in the H1 and V3 phases. The $\pm 1''$ is such a tiny offset that it would have required stimulus images of much higher resolution to render; we substituted it with a $5''$ vernier to retain consistency with other simulations and to reduce computational demands during parameter selection runs. The predicted performance on the small offset is only slightly more than the empirical performance in the $1''$ vernier. The bias control (reflecting the local response history) was implemented separately for vertical and horizontal tasks. In this experiment the horizontal and vertical tasks were in any event trained in different sessions, so it is reasonable for them to be separate.

Figure 5 shows the design of Herzog et al. (2006) experiment 3, the empirical results, and the corresponding predictions of the AHRM model. The model predictions are rank-order consistent with the observed data (Kendall's $\tau = 0.\ 575$, $p \ll 0.01$). The simulation shows a good qualitative fit of the data pattern: The bias developed for vertical and horizontal verniers is independent. The proportion variance accounted for by the model is $r^2 = 0.667$ ($p < 0.01$). Model parameters, selected to approximately parallel the data, are also listed in Table 1. A time-intensive grid search of the model parameters might improve the quantitative fit, which tracks the sum of the squared deviations between the model predictions and the data. In this case the model and data also differ for structural reasons. Our simulation did not implement the choice of Herzog et al. (2006) to assign the direction of induced bias in initial training in the same direction as a predetermined response bias for each observer, which led to an amplified initial bias in the empirical data. The simulation shows the predicted results for symmetric initial weights (priors) and a criterion control unit seeking a 50%–50% response distribution. The AHRM could be modified to incorporate either initial or ongoing preference for one response over the other in criterion control to mimic a natural bias in responses; we elected not to complicate the simulation in this way. Instead, we focus here on predicting the striking general patterns in the data in which the responses to the larger left and right stimuli diverge during the false feedback phase and converge once feedback is corrected.

The AHRM accounts for the oppositely induced biases from the feedback to train the weights for the channels around the circular orientation dimension. The AHRM representation units were extended to 12 orientations to cover both vertical and horizontal stimuli. Activations for the two sets of offsets (V and H) are focused in separate orientation (and spatial frequency) tuned representation units, so learned biases that are incorporated in learned weights are easily segregated in the model. This is essentially specificity arrived at by segregation in the representation domain. See Appendix B for examples and further discussion of the weight change profiles for this experiment.

Opposite induced biases in different spatial locations

The next demonstration of specificity showed that induced biases could be trained in opposite directions in different spatial locations of the display for vertical vernier judgments. Herzog et al. (2006) experiment 2 used a simplified induction design with only a single larger paired offset (i.e., $\pm 15''$) and one smaller offset (i.e., $-5''$ for the left location and $+5''$ for the right location) that received reverse (false) feedback. This experiment created opposite induced biases in the two locations on the screen.

Figure 6 illustrates the stimuli (Figure 6a) and the empirical data (Figure 6c). A single two-line vernier test in one of the two locations is trained on each trial. To model perceptual learning in different spatial locations, we used the IRT (Dosher et al., 2013; see Figure 6b for a schematic illustration). The theory was designed on principles similar to those of the AHRM but uses multilevel or multilayer representation structure to make predictions about transfer and specificity over learning in different spatial locations. It includes both location-specific representation layers and a location-independent representation layer. The model makes predictions about learning in several locations trained in interleaved training protocols, as in this experiment. Figure 6d shows predictions of the IRT for the spatially separate bias experiment. Note that this graph uses the format and labels of the source article, which labels the two large offset stimuli as *singleton* and *partner (large)* and the smaller offset stimuli as *partner (small)*; the graph shows the hit rate for each (rather than percentage *right*). The induced biases are in the opposite direction in the two locations.

The IRT framework makes predictions that are rank-order consistent with the observed data (Kendall's $\tau = 0.750$, $p \ll 0.01$). The percentage variance accounted for by the model is $r^2 = 0.860$ ($p < 0.01$). Model parameters were selected to approximate the levels in observed data (see Table 1). Further search of the parameter space might improve the match to overall performance level and hence the proportion of variance accounted for by the model. The model is exactly symmetric in the two locations, while the data appear to show lower performance for the right location (Figure 6c, right), accounting for much of the reduction in the proportion of variance accounted for by the model. This might have been handled by

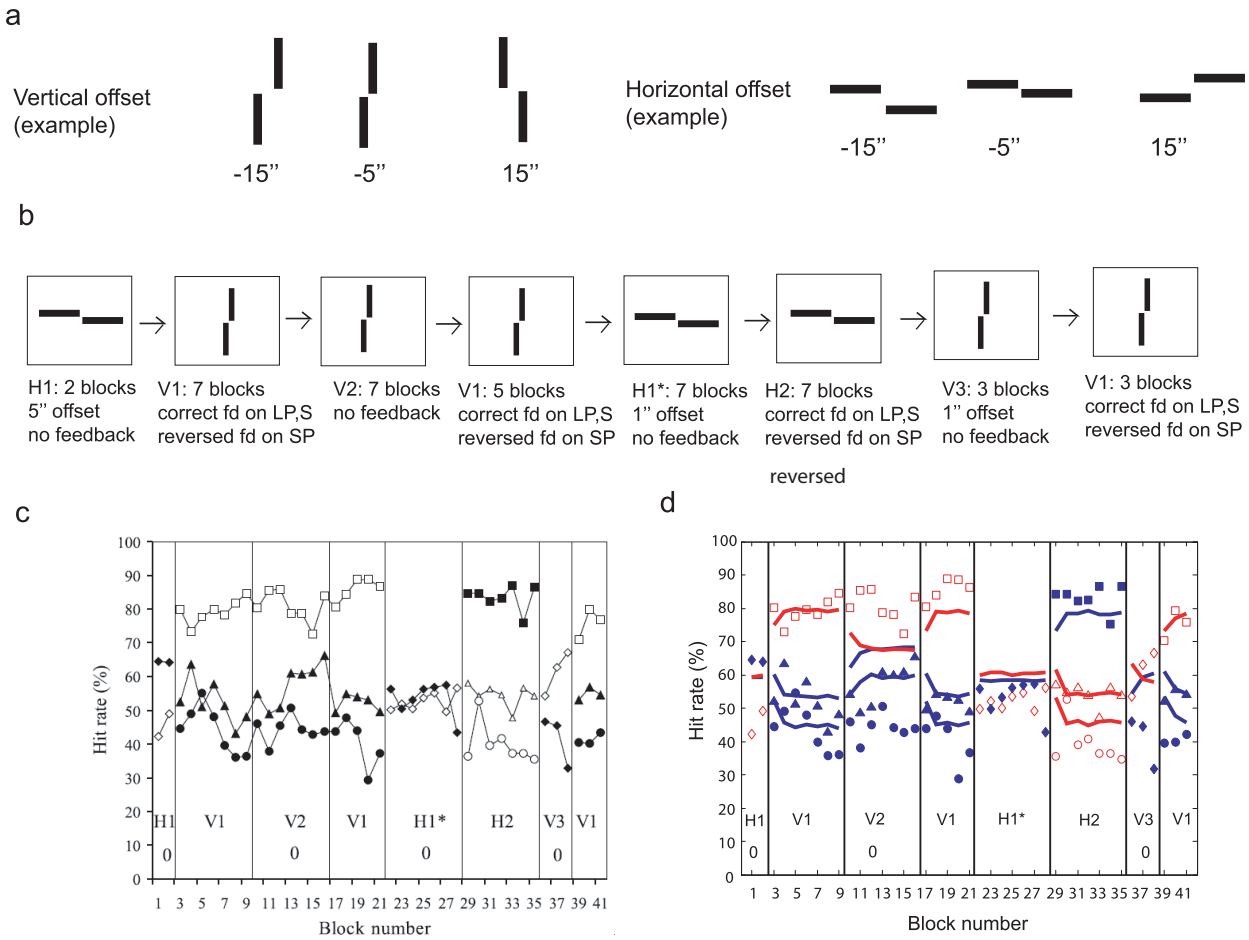

Figure 5. Separate and opposite biases induced by opposite feedback training for horizontal and vertical offset verniers and predictions of the AHRM. (a) The three vernier stimuli and reverse feedback regime. (b) The illustration of training phases and number of training blocks. In order, phase H1 tested small ($\pm 5''$) horizontal offsets, phase V1 trained vertical offsets ($-15''$, $-5''$, $+15''$) with reverse feedback for the small ($-5''$) offset to induce bias in the direction of the natural response bias, phase V2 tested V1 stimuli without feedback, phase V1 was repeated, phase H1* tested very small ($\pm 1''$) horizontal offsets without feedback, phase H2 trained horizontal offsets ($-15''$, $+5''$, $+15''$) with reverse feedback to induce an opposite bias to V1, phase V3 tested very small ($\pm 1''$) vertical offsets without feedback, and V1 was repeated. (c) Observed hit rate across these phases of training from Herzog et al. (2006, figure 7). (d) Predictions of the AHRM model for the conditions in panel c, using $5''$ instead of $1''$ in H1 and V3. Lines are simulations, and symbols are data. Panels b and c are adapted with permission from Herzog, M.H., Eward, K.R.F., Hermens, F., & Fahle, M. (2006). “Reverse feedback induces position and orientation specific changes.” *Vision Research*, *46*(22), 3761–3770.

incorporating differences in scaling factors or internal noises in the two locations, but we elected not to pursue this because, as indicated by the high Kendall's τ, the model does a very good job of accounting for the qualitative patterns of opposite bias induction followed by convergence with accurate feedback. The biases developed for two locations are generally independent of each other in both the data and the model.

In the IRT, location-independent representations are balanced over trials in reverse feedback for *right* and *left* responses, and the biases induced from reversed feedback carried average out in the weights from the location-independent representation layer to decision, so this layer does not contribute to the induced biases. The running average of postsynaptic activation of verniers was independently tracked for each location, which supported segregated bias learning. Appendix B provides examples of changing model weights and discusses the impact of location-specific representations and other implementation details.

Discussion

Herzog and colleagues demonstrated a very interesting series of phenomena related to induced bias, as well as improved discrimination, that can result from perceptual training. These training conditions pose challenges for models of perceptual learning because

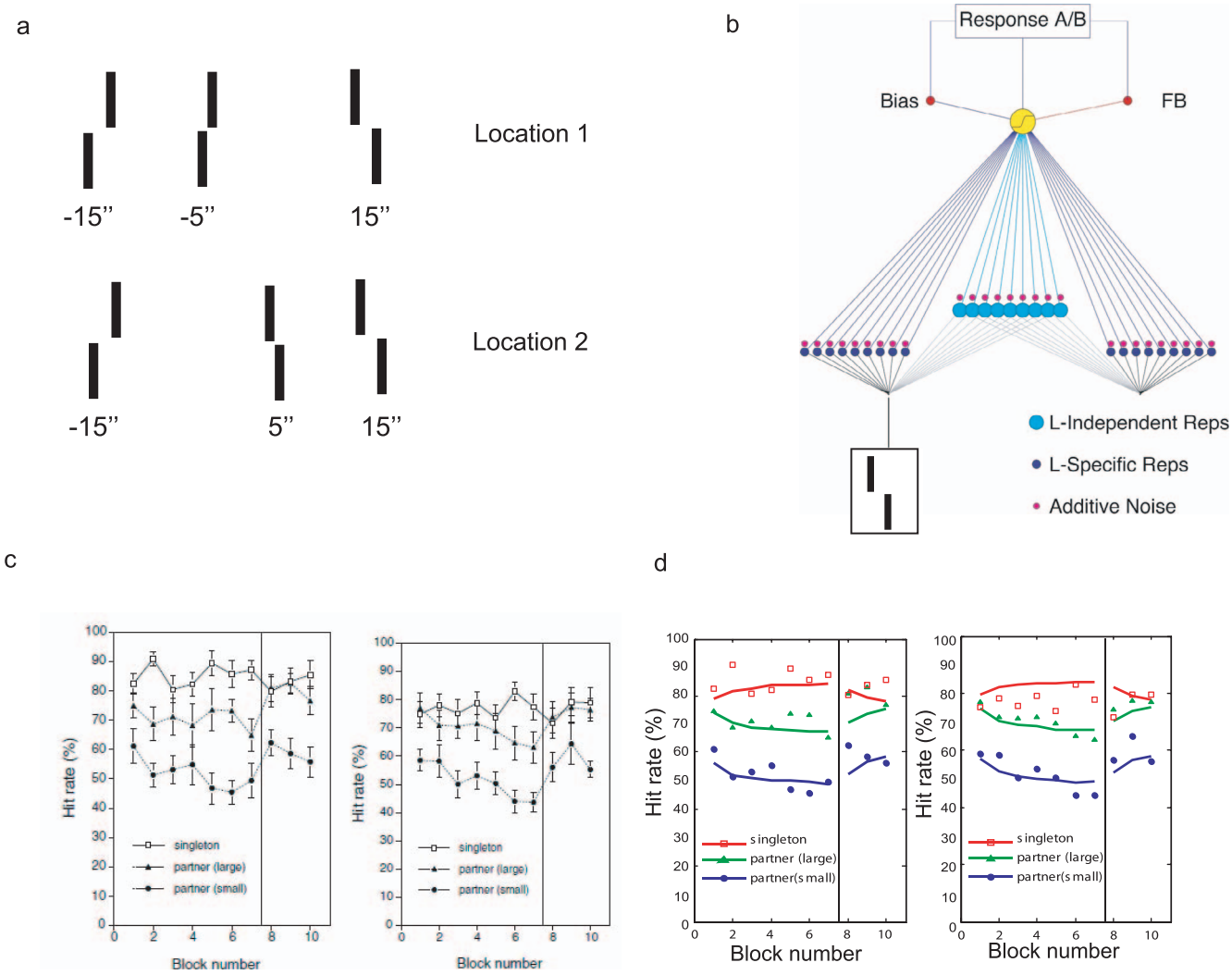

Figure 6. Inducing opposite biases in different spatial locations and predictions of the IRT, which accounts for perceptual learning in different spatial locations. (a) Opposite asymmetric stimulus sets and opposite reverse feedback used to induce opposing biases in the two spatial locations. A single two-line offset in one is trained on each trial. (b) The IRT schema extends the AHRM to multiple layers of stimulus representation, including location-specific and location-independent representations. (c) Data from Herzog et al. (2006, figure 5). (d) Predictions from IRT in the conditions in panel c in lines and data in symbols. Panel c is reproduced with permission from Herzog, M.H., Eward, K.R.F., Hermens, F., & Fahle, M. (2006). “Reverse feedback induces position and orientation specific changes.” *Vision Research*, *46*(22), 3761–3770.

they extend the generality of these models to situations that provide faulty feedback and/or imbalanced training sets. Both may be relevant to training situations in the real world.

In this article, we have demonstrated that the AHRM (Liu et al., 2010, 2012, 2014; Lu et al., 2010; Petrov et al., 2005, 2006) and the IRT (Dosher et al., 2013) that extends reweighting to multiple locations and multilevel representations account for the major phenomena in induced bias in the target articles, including the following: (a) bias induction by reverse feedback in asymmetric training sets; (b) the patterns of induced bias with different feedback and training regimens, including the importance of trial-by-trial feedback for achieving the bias induction result; (c) the ability to induce separate and opposite biases for sufficiently different stimuli that activate separate parts of the representation space; and (d) the ability to induce opposite biases for tasks in different spatial locations. While there may be subtle aspects of performance that the model in its current implementation only approximates, the perceptual learning by reweighting framework provides an excellent account of the broad phenomena of false or reversed feedback and induced biases in performance.

Here, we showed the consistency of the model by documenting that the model predicts qualitative patterns that parallel those documented in the corre-

sponding experimental paradigms. The qualitative and ordinal predictions of the model characterize most of the relevant regions of the parameter space of the models. Further (prohibitively) time-consuming fits might improve the quantitative fit of the model to data—largely by finding the exact interacting combination of internal noises and nonlinearities in the decision summations. Or, exact fits might require slightly different implementations of stimulus representations or elaboration of the representation layers to simulate statistical populations of tuned units.

All these phenomena of induced bias are consistent with (a) the theoretical framework that models perceptual learning as predominantly accomplished by incremental reweighting of evidence from stable sensory representations and (b) the broad principles of augmented Hebbian learning, which incorporates feedback when available. Recent work of others has suggested combined models of perceptual learning in which sensory representations are altered in some cases, whereas reweighting dominates perceptual learning in others (Sasaki et al., 2010). While we cannot rule out changes in the sensory representations through training, neither are they necessary to account for these phenomena of induced bias. Indeed, the fact that induced bias is so often reversible in relatively few sessions seems most compatible with trained changes in the learned weights between representation and decision units, or properties of session-specific bias tracking and control. Of these two, the observed induced biases are accomplished in the model by changing the learned weights in the direction of the feedback, while input from the bias control unit may serve to moderate these learned effects.

Changing model weights shifts the evidence distribution, and both shifting evidence and changing criterion contribute to the d' and criterion c estimates of standard signal detection. Indeed, the model predicts that the learned biases predominantly reflect shifts in evidence distributions feeding into decision—and only secondarily as compensatory variations in criterion offsets. This same caveat is true for all signal detection theory–based estimates in evaluation of behavioral data: What looks like shift in bias or criterion can in many cases be equivalently produced by a shift in evidence distributions. That is, moving a criterion down can be equivalent to moving the mean of evidence distributions up.

The augmented Hebbian learning system of the AHRM and IRT accommodates both unsupervised and supervised learning: It allows learning in the absence of feedback while incorporating available feedback. The effect of trial-by-trial feedback is obviously critical in the induced bias paradigms. Other models of perceptual learning that are broadly consistent with the reweighting framework but that use different learning mechanisms include an adaptive precision pooling model (Jacobs, 2009) and reward-based learning (Law & Gold, 2009). The adaptive precision pooling model, using the orientation–spatial frequency visual front end of the AHRM but Bayesian optimal estimation, makes predictions about learning that largely exceed behavioral performance; it is unclear how well it would account for the bias induction data. The architecture of the reward model (Law & Gold, 2009) is essentially parallel to that of the AHRM, but the representations were developed for motion discrimination. Detailed computational evaluation of these alternative learning mechanisms would require development and testing essentially comparable to the tests provided for the AHRM and IRT models that use the augmented Hebbian learning mechanism.

The AHRM and IRT models here predict some second-order effects for specific successive training regimens (i.e., potential effects of interleaving no-feedback training between reversed-feedback and accurate-feedback cycles; see Appendix B) that would require new experiments with other training manipulations to test or that, if the predictions are not verified, might suggest technical modifications in the implementation of Hebbian learning. While interesting, these are beyond the scope of the current article.

In these designs, stimulus frequencies (and feedback) were biased. Observers, however, likely have an a priori expectation preferring equal response frequencies. The current simulation approximated the observed response proportions in the empirical data quite well with a modest weight on criterion control tracking deviations from balanced responding. In previous simulations of trial-by-trial feedback—the only cases where an unbalanced situation may be obvious to the observer—feedback dominates over the criterion control in learning (Liu et al., 2014). Although a criterion control unit tracking a known unequal stimulus frequency may be more optimal in some situations, our data did not require it. Furthermore, absent explicit instructions, one would need a theory of how observers settled on a particular unequal response target. There are several possible implementations of this, and each would require significant investigation.

Aberg and Herzog (2012) also discuss factors that stand outside of the domain of the AHRM and IRT models as currently implemented. They argue that sensitivity and bias criterion effects may be differentially sensitive to overnight consolidation, based on the similarity of the derived criteria for reverse and correct trial-by-trial feedback conditions in block 11, immediately following a night of sleep. Our computational model framework does not currently incorporate specific mechanisms for overnight consolidation, although the postsynaptic running average $\bar{o}$ (see Appendix A) is reset in a new session. The AHRM

model seems to provide reasonable predictions at the level of performance accuracy, and the variation in criterion estimates focus on the data for a single block after an overnight sleep. Additionally, effects of overnight consolidation (Karni, Tanne, Rubenstein, Askenasy, & Sagi, 1994) are not always observed (Aberg, Tartaglia, & Herzog, 2009); such effects may occur more prominently in certain tasks. Other factors outside the usual models of perceptual learning (Herzog & Fahle, 1998) may also require extensions of the model framework.

In summary, this article has demonstrated the ability of a reweighting framework in general, and the AHRM and IRT models in particular, to broadly account for the interesting effects of recent induced bias paradigms in perceptual learning. In the model, the induced biases are predominantly encoded in learned weight structures that connect stable sensory representations of the stimuli to decision structures. Further computational investigation may elucidate whether other similar computational models with somewhat different implementations of stimulus encoding, decision, or learning modules are similarly able to account for these data. The current article extended the application range of the reweighting models to training situations with asymmetric sets of training stimuli and therefore asymmetric feedback ratios associated with reverse or correct feedback.

Keywords: perceptual learning, reverse feedback, asymmetric training, augmented Hebbian learning, bias

Acknowledgments

This research was supported by Grant EY-017491 from the National Eye Institute of the National Institutes of Health.

Commercial relationships: none.
Corresponding author: Barbara A. Dosher.
Email: bdosher@uci.edu.
Address: Department of Cognitive Sciences, University of California, Irvine, CA, USA.

References

Aberg, K. C., & Herzog, M. H. (2012). Different types of feedback change decision criterion and sensitivity differently in perceptual learning. *Journal of Vision, 12*(3):3, 1–11, doi:10.1167/12.3.3. [PubMed] [Article]

Aberg, K. C., Tartaglia, E. M., & Herzog, M. H. (2009). Perceptual learning with Chevrons requires a minimal number of trials, transfers to untrained directions, but does not require sleep. *Vision Research, 49,* 2087–2094.

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

Dosher, B. A., Chu, W., Liu, J., & Lu, Z.-L. (2012). Perceptual learning of task mixtures. *Journal of Vision, 12*(9): 767, doi:10.1167/12.9.767. [Abstract]

Dosher, B. A., Jeter, P., Liu, J., & Lu, Z.-L. (2013). An integrated reweighting theory of perceptual learning. *Proceedings of the National Academy of Sciences, USA, 110,* 13678–13683.

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

Dosher, B. A., & Lu, Z.-L. (1999). Mechanisms of perceptual learning. *Vision Research, 39,* 3197–3221.

Dosher, B. A., & Lu, Z.-L. (2009). Hebbian reweighting on stable representations in perceptual learning. *Learning & Perception, 1,* 37–58.

Fahle, M., & Poggio, T. (2002). *Perceptual learning*. Cambridge, MA: MIT Press.

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

Herzog, M. H., Eward, K. R. F., Hermens, F., & Fahle, M. (2006). Reverse feedback induces position and orientation specific changes. *Vision Research, 46,* 3761–3770.

Herzog, M. H., & Fahle, M. (1997). The role of feedback in learning a vernier discrimination task. *Vision Research, 37,* 2133–2141.

Herzog, M. H., & Fahle, M. (1998). Modeling perceptual learning: Difficulties and how they can be overcome. *Biological Cybernetics, 78,* 107–117.

Herzog, M. H., & Fahle, M. (1999). Effects of biased feedback on learning and deciding in a vernier discrimination task. *Vision Research, 39,* 4232–4243.

Huang, C. B., Lu, Z.-L., & Dosher, B. A. (2012). Co-learning analysis of two perceptual learning tasks with identical input stimuli supports the reweighting hypothesis. *Vision Research, 61,* 25–32.

Jacobs, R. A. (2009). Adaptive precision pooling of model neuron activities predicts the efficiency of human visual learning. *Journal of Vision, 9*(4):22, 1–15, doi:10.1167/9.4.22. [PubMed] [Article]

Jones, P. R., Moore, D. R., Amitay, S., & Shub, D. E. (2013). Reduction of internal noise in auditory perceptual learning. *The Journal of the Acoustical Society of America, 133,* 970–981.

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J., & Sagi, D. (1994). Dependence on REM sleep of overnight improvement of a perceptual skill. *Science, 265,* 679–682.

Kendall, M. G. (1938). The new measure of rank correlation. *Biometrika, 30,* 81–93.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). London, United Kingdom: Griffin.

Law, C. T., & Gold, J. I. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature Neuroscience, 12,* 655–663.

Liu, J., Dosher, B., & Lu, Z.-L. (2014). Modeling trial by trial and block feedback in perceptual learning. *Vision Research, 99,* 46–56.

Liu, J., Lu, Z.-L., & Dosher, B. A. (2010). Augmented Hebbian reweighting: Interactions between feedback and training accuracy in perceptual learning. *Journal of Vision, 10*(10):29, 1–14, doi:10.1167/10.10.29. [PubMed] [Article]

Liu, J., Lu, Z.-L., & Dosher, B. A. (2012). Mixture of training at high and low accuracy levels facilitates learning at low training accuracy level. *Vision Research, 61,* 15–24.

Lu, Z.-L., & Dosher, B. A. (2012). Visual perceptual learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 3415–3418). Berlin, Germany: Springer.

Lu, Z.-L., Liu, J., & Dosher, B. A. (2010). Modeling mechanisms of perceptual learning with augmented Hebbian re-weighting. *Vision Research, 50,* 375–390.

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal, 2,* 45–64.

Petrov, A., Dosher, B. A., & Lu, Z.-L. (2005). Perceptual learning through incremental channel reweighting. *Psychological Review, 112,* 715–743.

Petrov, A. A., Dosher, B. A., & Lu, Z. L. (2006). Perceptual learning without feedback in non-stationary contexts: Data and model. *Vision Research, 46,* 3177–3197.

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science, 256,* 1018–1021.

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

Sagi, D. (2011). Perceptual learning in vision research. *Vision Research, 51,* 1552–1566.

Sasaki, Y., Nanez, J., & Watanabe, T. (2010). Advances in visual perceptual learning and plasticity. *Nature Reviews Neuroscience, 11,* 53–60.

Schoups, A. A., Vogels, R., & Orban, G. A. (1995). Human perceptual learning in identifying the oblique orientation: Retinotopy, orientation specificity and monocularity. *The Journal of Physiology, 483*(Pt. 3), 797–810.

Tlapale, E., Dosher, B. A., & Lu, Z.-L. (2013). Modeling perceptual learning of visual motion. *Journal of Vision, 13*(9): 248, doi:10.1167/13.9.248. [Abstract]

Tlapale, E., Dosher, B. A., & Lu, Z.-L. (2014). Modeling motion perception in random dot kinematograms. *Journal of Vision, 14*(10): 474, doi:10.1167/14.10.474. [Abstract]

Weiss, Y., Edelman, S., & Fahle, M. (1993). Models of perceptual learning in vernier hyperacuity. *Neural Computation, 5,* 694–718.

Wenger, M. J., & Rasche, C. (2006). Perceptual learning in contrast detection: Presence and cost of shifts in response criteria. *Psychonomic Bulletin & Review, 13,* 656–661.

Appendix A

This article examines the proposition that perceptual learning of sensitivity and bias is consistent with incremental reweighting of sensory evidence (Dosher & Lu, 1998, 1999). The AHRM simulates a multichannel network model. It takes stimulus images as input, produces a task response, and reweights (updates weights) the connections from representation units to a decision unit. Learning is augmented by inputs from feedback and from a criterion control unit. Appendix A provides a brief summary of representation, decision, and learning subsystems or modules of the corresponding implementation of the AHRM, shown schematically in Figure 2. Descriptions of the model can be found in previous studies (Liu et al., 2010, 2012, 2014; Lu et al., 2010; Petrov et al., 2005, 2006). A related theoretical framework, the IRT (Dosher et al., 2013), is briefly discussed at the end of Appendix A.

The representation subsystem (module) used in this article consists of orientation- and frequency-selective units. The orientation evidence coded in the activations of these orientation–spatial frequency units is used to discriminate the "tilt" of the line offset vernier stimuli.

This system has been used to model perceptual learning performance for tasks involving the discrimination of the orientation of Gabor patches (i.e., Petrov et al., 2005) and of depth tilt of grid patterns (Jacobs, 2009). Alternative representation modules have been used to model learning in motion tasks (Tlapale et al., 2013, 2014) or three-point vernier or bisection judgments (Huang et al., 2012).

The activation values of the orientation–spatial frequency filters $A(\theta,f)$ compute the normalized spectral energy in the image in each channel. First, retinotopic phase-sensitive maps $S(x,y,\theta,f,\phi)$ are computed for the input image $I(x,y)$:

$$S(x,y,\theta,f,\phi) = \left[RF_{\theta,f,\phi}(x,y) \otimes I(x,y)\right]^2. \quad (1)$$

These units at location (x,y) are tuned to spatial frequency f, orientation θ, and spatial phase ϕ. The set of filters consisted of the joint product of five spatial frequencies (8, 11.3, 16, 22.6, 32 cycles/degree), seven orientations (0°, ±15°, ±30°, ±45°), and four spatial phases (0°, 90°, 180°, 270°). (The simulation with both vertical and horizontal verniers requires 12 orientations spanning from −90° to 75° with a step size of 15°.) Spatial frequency tuning and orientation tuning bandwidths were set at $h_f = 1$ octaves and $h_\theta = 30°$ (half amplitude, full bandwidth). These values are the same as those used in prior applications of this form of the AHRM (Dosher et al., 2013; Liu et al., 2010, 2012; Petrov et al., 2005, 2006) and were based on estimates of cellular tuning bandwidths in the primary visual cortex.

Using the fast Fourier transform, the input image $I(x, y)$ is convolved with each unit, followed in succession by half-squaring rectification, spatial phase pooling, and then inhibitory normalization (Heeger, 1992), respectively:

$$E(x,y,\theta,f) = \sum_{\phi} S(x,y,\theta,f,\phi) + \varepsilon_1 \quad (2)$$

and

$$C(x,y,\theta,f) = \frac{aE(x,y,\theta,f)}{k + N(f)}. \quad (3)$$

In these equations, the normalization pool N_f is tuned weakly for spatial frequency and is independent of orientation (see Petrov et al., 2005). a is a scaling factor; the saturation constant k is relevant for extremely small contrasts. In this application, we pool over spatial phase and a stimulus evidence region with kernel of radius W_r.

The representation activations become stochastic due to two internal noises. The internal additive noise term ε_1 has mean 0 and standard deviation σ_1, with a Gaussian distribution. The internal multiplicative noise ε_2 of mean 0 and standard deviation σ_2 introduces another source of stochastic variability. The activation in each orientation and spatial frequency tuned unit is computed as follows:

$$A'(\theta,f) = \sum_{x,y} W_r(x,y)C(x,y,\theta,f) + \varepsilon_2 \quad (4)$$

This intermediate value is passed through an activation function with gain parameter γ that range limits the final activation of the representation units:

$$A(\theta,f) = \begin{cases} \dfrac{1 - e^{-\gamma A'}}{1 + e^{-\gamma A'}} A_{\max}, & if A' \geq 0 \\ 0, & otherwise \end{cases}. \quad (5)$$

The decision subsystem combines the activation pattern over the representation units to yield a decision by weighting the input from each representation unit (35 units for AHRM, 60 for IRT) by w_i and a bias factor b with weight w_b and incorporating random Gaussian decision noise ε_d (mean 0 and standard deviation σ_d):

$$u = \sum_{i=1}^{35} w_i A(\theta_i,f_i) - w_b b + \varepsilon_d. \quad (6)$$

The "early" activation of the decision unit o' is a sigmoid function of the weighted activations u with gain γ:

$$o' = G(u) = \frac{1 - e^{-\gamma u}}{1 + e^{-\gamma u}} A_{\max} \quad (7)$$

A negative o' maps to one response (*left*), while a positive o' maps to the other response (*right*).

The learning subsystem updates the synaptic connection weights from sensory representation units to the decision unit on every trial. If trial-by-trial feedback ($F = \pm 1$) is available, it shifts the activation in the decision unit to a late level o:

$$o = G(u + w_f F) \ \ (late). \quad (8)$$

Hebbian learning processes operate on the late phase activation of the decision unit. If trial-by-trial feedback is available and the feedback weight is relatively high, then the activation will go to its maximum ($\pm A_{\max} = \pm 1$); smaller feedback weights may only slightly shift activation toward the correct response. If feedback is not present, learning operates without the benefit of this shift toward a correct response ($o = o'$). Except for very low accuracy conditions, the learned weights tend to move toward a more optimal weight distribution because o' tends to correlate with the correct response.

The amount of change in connection weights depends jointly on the learning rate η, the presynaptic activation $A(\theta,f)$, how far the postsynaptic activation is

from its long-term average, $(o - \bar{o})$, and the distance between the current weights and their saturation values, $w_{\min}$ or $w_{\max}$. Weights are changed (learned) according to this rule:

$$\Delta w_i = (w_i - w_{\min})[\delta_i]_- + (w_{\max} - w_i)[\delta_i]_+ \quad (9)$$

with

$$\delta_i = \eta A(\theta_i, f_i)(o - \bar{o}) \quad (10)$$

and average postsynaptic activation of

$$\bar{o}(t+1) = \rho o(t) + (1-\rho)\bar{o}(t). \quad (11)$$

The average postsynaptic activation $\bar{o}$ is inherited from block to block and is independently tracked for inputs of different orientations or from different locations. It is only reset between sessions, such as after a night's sleep as in Aberg and Herzog (2012). This treatment better replicates the qualitative pattern of the behavioral results.

The role and implementation of the top-down bias control unit is to balance the ongoing frequency of the *left–right* decisions, which indirectly augments learning. The bias correction term b tracks deviations of the recent response frequencies from 50% (or the instructed presentation probabilities) of the simulated observer. Criterion control input b weighted by w_b is input to the decision unit. The bias on each trial is an exponentially weighted average of the responses with a time constant of 50 trials ($\rho = 0.02$):

$$r(t+1) = \rho R(t) + (1-\rho)r(t), \quad (12)$$

$$b(t+1) = \alpha r(t). \quad (13)$$

$R(t)$ is the current trial's response (*left* = −1 and *right* = +1), and $r(t)$ is the response running average that exponentially discounts past trials. Bias control is more important to learning in the absence of trial-by-trial external feedback (Petrov et al., 2006).

The bias correction term shifts the response criterion to counterbalance a shift in the proportion of *right* responses. The experiments by Herzog and colleagues present more *left* stimuli but systematically bias feedback toward *right* and are designed to generate biases in responding. This is equivalent to shifting the criterion in a compensatory direction. Higher bias weights (w_b) increase the impact of the bias correction term. Liu et al. (2014) used a hypothesized relationship between the accuracy in the last block of trials—either from block feedback or estimated in trial feedback conditions—and the bias weight (w_b). In essence, the system has more confidence in the bias information when accuracy is high and less confidence in the bias information when accuracy is low. The minimum and maximum of the bias weight are at 0 and 1 for performance accuracies (proportion correct) between chance at 0.50 and perfect performance at 1.0, with the bias weight set to twice the percentage correct minus one. The bias weight changes after every block in the block-feedback conditions.

Appendix B

Appendix B presents and discusses aspects of the learned weight structures in the AHRM or IRT models for the model fits to experiments provided in the article. The vernier stimuli are so similar to one another that most of the reliable information is carried in changes in the activity levels of orientation channels very near the vertical (or horizontal for horizontal vernier judgments). The AHRM and IRT models begin with weights that build in prior knowledge of left and right tilting patterns in vertical vernier judgments: $w_i = (\theta / 45)w_{init}$. Learning and bias in these experiments reflect relatively subtle changes that tilt these weights toward one response or the other; the *left* and *right* vernier is very similar in representation space (see Figure 3b), and the percentage changes of weights are in many cases quite small. In order to make these subtle changes more visible, we display changes in the weights as a function of training, relative to initial values. We chose to scale these as proportional changes relative to the average magnitude of all the initial weights. In previous applications of the model to experiments with widely varying stimuli, changes were visible in the weights themselves (Dosher et al., 2013; Liu et al., 2014; Petrov et al., 2005, 2006).

Inducing bias with asymmetric training and false feedback

The simulation of Herzog and Fahle (1999) experiment 3 modeled the primary induction paradigm. Percentage changes in the weights from different orientation and spatial frequency channels are shown for the successive phases of reverse trial-by-trial feedback and accurate feedback training (after the vertical line) in Figure B1. The color of the line codes the orientation of the channels, and lines of the same color are for different spatial frequencies. Training with reverse feedback on the smallest left stimulus shifts the weights upward, tracking the dominance of *right* feedback. Subsequent training with accurate feedback shifts weights toward left, now tracking the more likely *left* feedback. Indeed, the largest shifts are for orientation channels of 0°, +15°, and −15° that are most sensitive to the very small angles of the vernier stimuli. The shifts in the first phase change more slowly than

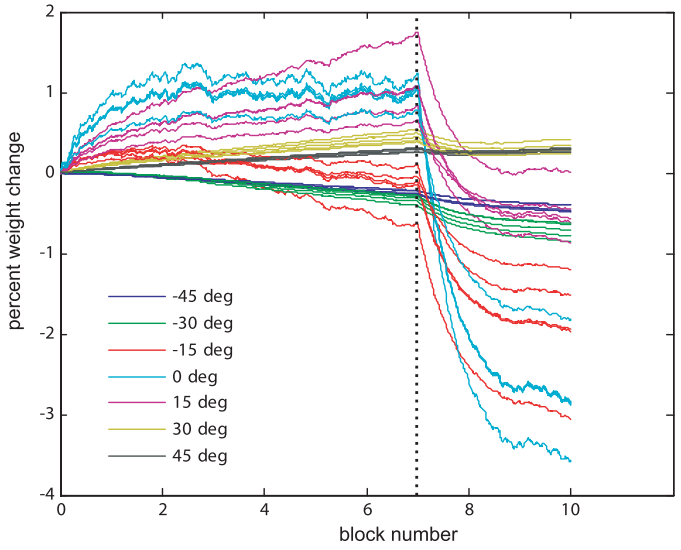

Figure B1. Percentage weight changes for different orientation and spatial frequency channels in the two phases of training for the model of Herzog and Fahle (1999, experiment 3). The color represents the orientation of the channels; multiple lines of the same color are for different spatial frequencies. Upward shifts in weights cause response shifts toward *right*, whereas downward shifts in weights cause response shifts toward *left*.

those after the switch to accurate feedback. This occurs for two reasons. In reverse feedback, the slight stimulus information in the small subthreshold offset in the initial weights opposes its false feedback, while with accurate feedback they move in the same direction. Additionally, in the Hebbian rule, the size of the weight change $\delta_i = \eta A(\theta_i f_i)(o - \bar{o})$ is proportional to the difference of the postsynaptic output and its average over time. The contrasts with a right-shifted average postsynaptic output, $\bar{o}$, inherited from the reverse feedback phase, increases the effective weight change δ at the beginning of the correct trial-by-trial feedback phase.

Induced biases depend on trial-by-trial feedback

The next simulation examined the ability of the AHRM model to handle the data of Aberg and Herzog (2012) comparing different kinds of feedback. In these panels, the first phase (up to the first vertical line) corresponds with the feedback condition of the label (e.g., no feedback, reverse trial-by-trial feedback); the second phase has three blocks of no feedback; and the third phase has three blocks of accurate trial-by-trial

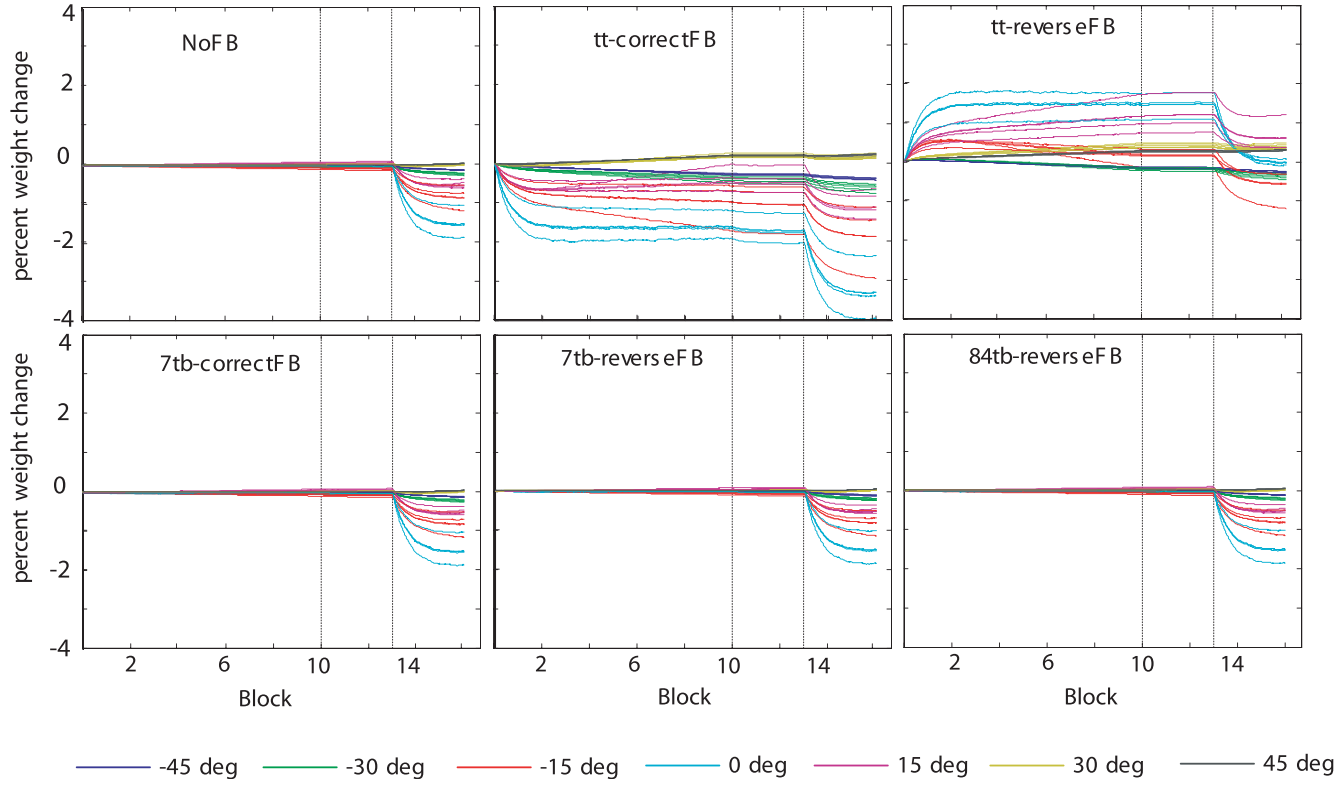

Figure B2. Percentage weight changes for different orientation and spatial frequency channels in the three phases of training of Aberg and Herzog (2012). The color conventions follow those in Figure B1. Training with reverse trial-by-trial feedback shifts the weights upward, corresponding to the bias to *right* feedback, whereas accurate trial-by-trial feedback shifts weights left, tracking the more frequent *left* stimuli and feedback.

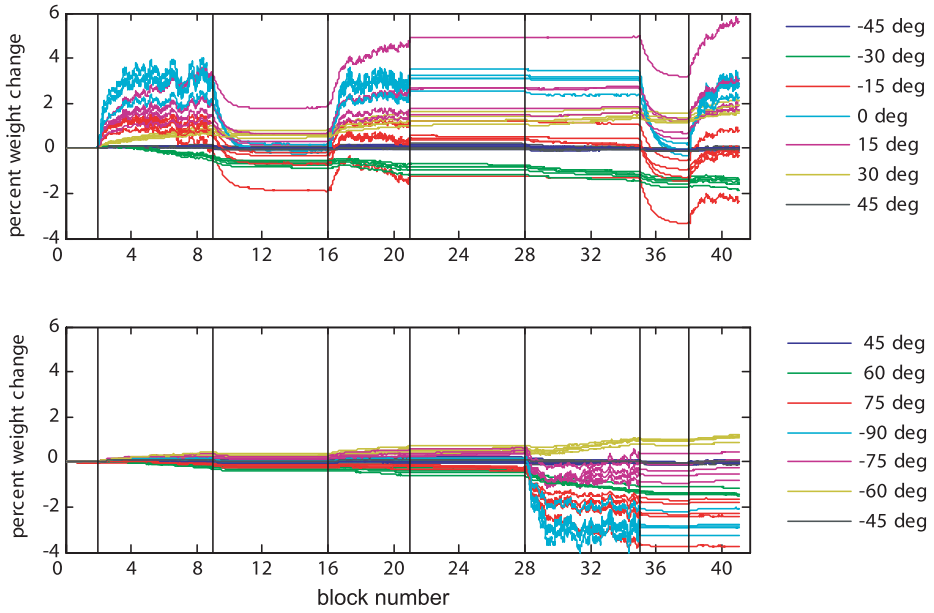

Figure B3. Percentage weight changes for different orientation and spatial frequency channels in the three phases of training of Herzog et al. (2006, experiment 3). The color conventions follow those of earlier figures. Orientations near vertical (top) and horizontal (bottom) are shown separately for clarity. Training with reverse trial-by-trial feedback shifts the weights upward for vertical stimuli and in the opposite direction for horizontal stimuli.

feedback. The patterns of weight change in the model are shown in Figure B2. Consistent with the pattern in the data and the predictions of the AHRM model, only the accurate trial-by-trial feedback or the reverse trial-by-trial feedback conditions substantially change performance in the first phase. Trial-by-trial feedback moves the postsynaptic output toward the biasing direction. Block feedback only changes the weight on the bias control unit, which operates to eliminate bias in the response frequencies. Correspondingly, these two trial-by-trial feedback conditions are the only ones that exhibit significant percentage weight changes in this phase of training. The no-feedback condition and all of the block-feedback variants show essentially no change. Any weight changes in the first phase are largely maintained during the middle three-block phase of practice without feedback. Then, the weights are shifted left, toward the dominant *left* feedback and the asymmetric stimuli in the correct (accurate) feedback in the final phase of training. The weight change in this third phase reprises that in the first phase of the correct-feedback condition, starting from the weight state at the end of phase two. This rerelease of new learning after the no-feedback training phase is also a peculiar interaction in which the size of weight change depends on the contrast of the postsynaptic output and its running average, or $(o - \bar{o})$. At the beginning of training, the running average begins at zero; as time goes on, the average postsynaptic activity trends negative, or *left*, and so the asymmetric left feedback has a smaller impact and weight change slows in the trial-by-trial feedback conditions. Three blocks of training with vernier stimuli without feedback, where the postsynaptic output reflects only stimulus information yielding postsynaptic outputs that are so close to zero, reinstates the conditions of early learning.

Although this prediction seems consistent with the marked down trend in percentage *right* data in the last three blocks of correct feedback training in the behavioral data (see Figure 6), it is not clear how strongly this feature of the model is tested in the current data sets. This peculiar predicted interaction with interspersed no-feedback training seems to be a property of the small offsets of the vernier stimuli combined with the asymmetric stimulus design. If taken seriously, this property seems to predict a possibly testable advantage to cycling feedback training with no-feedback training.

Independent (opposite) induced bias in different trained orientations

The AHRM model fairly naturally predicts the specificity of induced bias to training of vertical and horizontal vernier offsets in experiment 3 of Herzog et al. (2006). To simulate this experiment, the number of orientation channels is increased to cover the full circular orientation dimension. The percentage weight changes from the simulation of the many subphases of

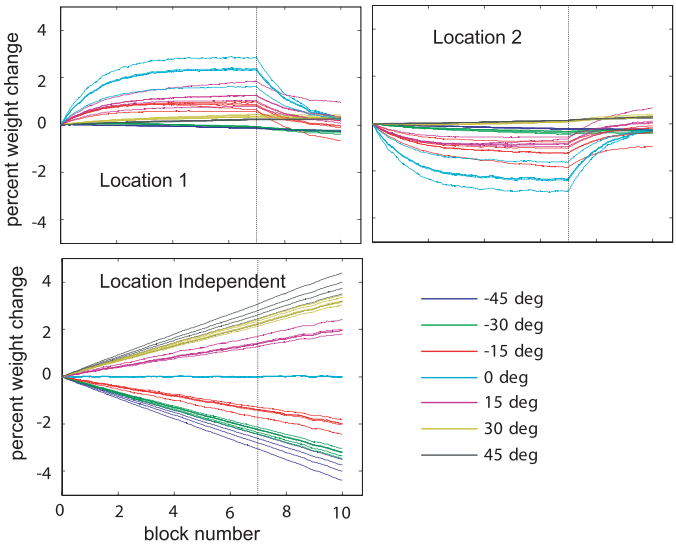

Figure B4. Percentage weight changes for different orientation and spatial frequency channels in location-specific and location-independent representations during the two phases of training of Herzog et al. (2006). The color conventions follow those of earlier figures.

training in this experiment are shown in separate panels for the orientations around vertical (top) and around horizontal (bottom) in Figure B3. In this design, the induced bias is in opposite directions. The induced bias for reverse feedback is scaled upward for vertical training (for consistency with the earlier graphs for the primary induction experiments); because the induced bias is in the opposite direction, it is scaled downward for horizontal training. In the experiment, these might correspond, for example, with right-hand and left-hand biases. The simulation tracks the average postsynaptic activity separately for horizontal and vertical judgments, which are instructed as different tasks and appear in different test or training blocks. As discussed in the Results, the model implementation approximated the $\pm 1''$ vernier offsets that were used essentially to measure pure bias in the system with $\pm 5''$ in order to retain smaller image representations.

After a brief baseline measure for small horizontal offsets without feedback, different variants of vertical vernier tasks were trained in succession, followed by near-zero horizontal vernier offsets without feedback beginning at block 21 and (opposite) reverse trial-by-trial feedback at block 35. Because the change in any given weight δ_i depends directly on the activity in that spatial frequency and orientation channel $A(\theta,f)$ and because that activity is largely focused on orientations near the vertical for vertical vernier judgments (and vice versa for horizontal), the weight changes on units relevant to horizontal judgments are largely unchanged (except for random drift) during the vertical vernier training phases. Similarly, the weight changes on units relevant to vertical judgments are largely unchanged (except for random drift) during the horizontal vernier training phases. Otherwise, the percentage weight changes during reverse feedback training and accurate feedback training phases mirror those described earlier.

Opposite induced biases in different spatial locations

Last, the IRT model was applied to the data from Herzog et al. (2006) experiment 2 showing that biases could be separately and oppositely induced for training in separate spatial locations. The IRT architecture and framework is an extension of the AHRM designed to account for transfer and learning interactions between training in separate spatial locations. Partial specificity to location is an often-reported property of perceptual learning (i.e., Schoups et al., 1995, for orientation). Learning the weights on location-specific sensory representations mediates location-specific perceptual learning, while learning the weights on location-independent representations scaffolds transfer from one location to another and accounts for interactions of training in different locations. Although the data might be consistent with a multilocation AHRM without the location-independent layer, we used the full IRT to simulate the experiment because it has been used previously to model other multilocation transfer (Dosher et al., 2013) or multilocation learning interactions (Dosher et al., 2012). The location-independent representations are more broadly tuned and noisier than the location-specific ones. In this simulation, the

average postsynaptic activities are tracked separately for the location-specific units.

The weight changes in the simulation that result from opposed asymmetric bias induction in the two locations are shown in Figure B4. The top panels show the percentage weight changes for the two separate location-specific representations, which show the now-familiar pattern for reverse trial-by-trial feedback bias induction (in opposite directions in the two locations) followed by a correction with a switch to correct (accurate) trial-by-trial feedback. The bottom panel shows steady and symmetric spread in the weights on the location-independent representations. These changes are somewhat faster, reflecting twice the number of total training trials; the symmetry reflects the cancellation of induced bias due to balanced trials in the left and right locations. Some of the improvements in discrimination performance are carried by the unbiased shifts in the location-independent weights, while the induced biases are carried by the location-specific weights. This is a demonstration that the IRT, which has been applied in other multilocation training paradigms, can model this opposed-bias experiment as well.

Summary

The AHRM simulations and the IRT simulation provide a consistent account of the basic bias-induction paradigm, the differential effects of different forms of feedback on performance, and the specificity of separate and opposite biases induced for vertical and horizontal vernier judgments and for vertical vernier judgments in separate spatial locations. Some specific attributes of the particular implementation of Hebbian learning, such as the role of the average postsynaptic activity levels in resetting more rapid learning in the transition between certain feedback regimes, might yield interesting, testable predictions.