Title
Learning spatial frequency identification through reweighted decoding.

Journal
Journal of vision, 23(6)

ISSN
1534-7362

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

Publication Date
2023-06-01

DOI
10.1167/jov.23.6.3

Peer reviewed

Learning spatial frequency identification through reweighted decoding

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

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

Zhong-Lin Lu Division of Arts and Sciences, NYU Shanghai, Shanghai, China; Center for Neural Science and Department of Psychology, New York University, NY, USA NYU-ECNU Institute of Brain and Cognitive Neuroscience, Shanghai, China

Perceptual learning, the improvement of perceptual judgments with practice, occurs in many visual tasks. There are, however, relatively fewer studies examining perceptual learning in spatial frequency judgments. In addition, perceptual learning has generally been studied in two-alternative tasks, occasionally in *n*-alternative tasks, and infrequently in identification. Recently, perceptual learning was found in an orientation identification task (eight-alternatives) and was well accounted for by a new identification integrated reweighting theory (I-IRT) (Liu et al., submitted). Here, we examined perceptual learning in a similar eight-alternative spatial frequency absolute identification task in two different training protocols, finding learning in the majority but not all observers. We fit the I-IRT to the spatial frequency learning data and discuss possible model explanations for variations in learning.

Introduction

Visual spatial patterns vary in many ways, but among the most salient features are their orientation and spatial frequency. Perceptual learning—the improvement in performance with practice or training—has been extensively studied in many visual tasks, including orientation (Crist, Kapadia, Westheimer, & Gilbert, 1997; Fiorentini & Berardi, 1997; Dosher & Lu, 1998; Dosher & Lu, 1999; Liu, 1999; Lu & Dosher, 2004), motion (Zhou et al., 2006), texture (Karni & Sagi, 1991; Ahissar & Hochstein, 1997), and hyperacuity (Poggio, Fahle, & Edelman, 1992; Fahle, Edelman & Poggio, 1995; Saarinen & Levi, 1995; Young, Li, Levi, Klein, & Huang, 2004; Fahle, 2005), usually in the context of two-alternative judgments. Less is known, however, about perceptual learning in the spatial frequency domain than in orientation or many other aspects of pattern stimuli. Similarly, perceptual learning in absolute identification tasks, or indeed in other *n*-alternative identification tasks, has been studied in only selected tasks. In identification tasks, a single stimulus is presented, and the observer classifies it into one of *n* responses. In absolute identification, stimuli vary along a single sensory dimension and there typically are four or more stimuli. In discrimination tasks, stimuli from all potential categories—usually two—are presented on each trial (either simultaneously or successively) and the observer chooses the response by the order of the presentation. In identification the observer must develop multiple internal stimulus representations (or multiple criteria), and the task may be more demanding. In contrast, discrimination tasks can directly compare the available stimuli, and so they are often thought to yield more precise fine discrimination performance (Stewart, Brown, & Chater, 2005). Given the lack of learning often reported in absolute identification (but see Rouder, Morey, Cowan, & Pealtz, 2004; Dodds, Donkin, Brown, & Heathcote, 2011) and the small number of studies about learning spatial frequency in any paradigm, the ability to learn absolute identification of spatial frequency is a comparatively open question.

Recently, we documented robust learning in an orientation absolute identification task with eight alternatives (Liu, Lu, & Dosher, *submitted*). We found that learning was strongly influenced by the nature

Citation: Dosher, B., Liu, J., & Lu, Z.-L. (2023). Learning spatial frequency identification through reweighted decoding. *Journal of Vision*, *23*(6):3, 1–31, https://doi.org/10.1167/jov.23.6.3.

https://doi.org/10.1167/jov.23.6.3 Received November 18, 2022; published June 2, 2023 ISSN 1534-7362

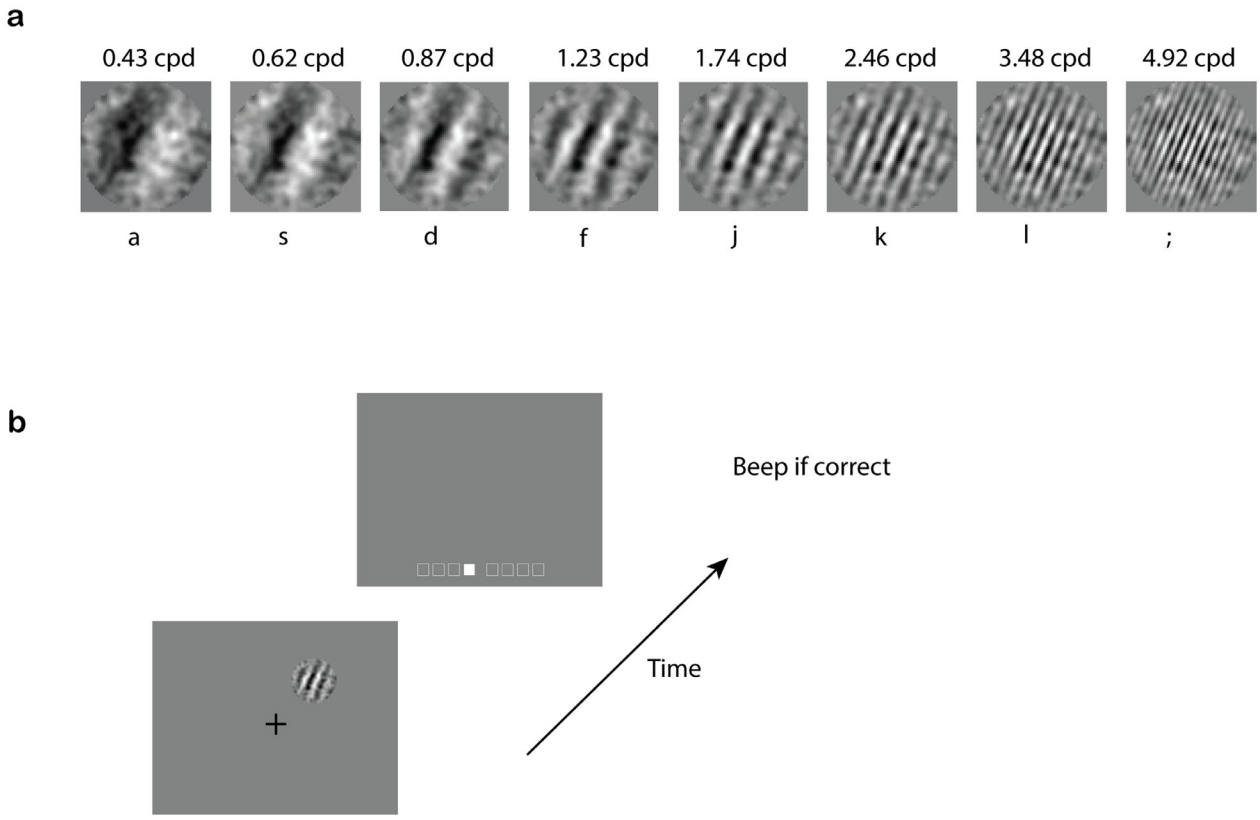

Figure 1. (**a**) The experimental stimuli – noisy Gabor images with eight possible spatial frequencies. (**b**) A simplified task paradigm: every trial only one stimulus was presented, and observers made a response. A visual feedback (white square) and audio feedback (beep if the response was correct) followed.

of feedback and that learning was well accounted for by a new identification integrated reweighting model (I-IRT), based on the original IRT model of perceptual learning in binary tasks (Dosher, Jeter, Liu, & Lu, 2013). The I-IRT learns through experience by improving the weights between stimulus representations and decision units – optimizing judgments through reweighting of evidence. The pattern of activations in stimulus representations is the encoding, and the weighting to decision is the decoding. Learning improves the decoding. Using this neural network framework, with the same early visual representations, the model predicts that perceptual learning should also occur for a corresponding absolute identification of spatial frequency. It is this prediction that we examine in this paper.

Perceptual learning in the spatial frequency identification task studied here (see Figure 1) touches on several literatures. We start by reviewing the previous examples of perceptual learning related to spatial frequency, the literature in visual perceptual learning in general, including multidimensional *n*-alternative identification, and finally the learning in absolute identification of unidimensional stimuli—including our prior study of learning in eight-alternative orientation identification, and predictions of the I-IRT framework. We then tested these predictions related to spatial frequency learning in two experiments.

Perceptual learning in spatial frequency tasks

Perceptual learning occurs in many tasks (see Fine & Jacobs, 2002; Sagi, 2011; Watanabe & Sasaki, 2015; Dosher & Lu, 2020; Lu & Dosher, 2022; for reviews), and, in some cases, learning of other judgments such as contrast discrimination has been specific to the spatial frequency of the stimuli (Yu, Klein, & Levi, 2004). Learning related to spatial frequency judgments has been examined in a few studies testing difference thresholds (Meinhardt, 2001; Meinhardt, 2002) and in several others identification or discrimination of compound spatial patterns have been studied (Fiorentini & Berardi, 1980; Fiorentini & Berardi, 1981; Fine & Jacobs, 2000). These studies demonstrated improved threshold differences in yes/no or two-interval binary comparisons, or improved binary discrimination using more complex compound stimuli (where learning

may have capitalized on changes in diagnostic spatial pattern features, such as changes in visible bands or in plaid stimuli).

In the earliest studies of perceptual learning of spatial frequency judgments, observers learned to identify two spatial patterns with varying spatial frequency components: a standard composed of 1f and 3f vertical sine wave components with 0 degree phase offset and lower contrast in the 3f component; observed learning was largely specific to orientation and viewing distance (Fiorentini & Berardi, 1980). A follow-up study showed learning to discriminate patterns in which the 3f component shifted phase or contrast, or by the addition of a higher 5f harmonic, using two-interval forced choice in which the observer selected the standard versus a variant in blocked tests (Fiorentini & Berardi, 1981). Discriminating the compound patterns improved over 200 trials, whereas discriminating frequency differences in simple gratings did not. A related study showed perceptual learning in a four-interval odd-out discrimination 5qwk for composite "wicker" or plaid pattern stimuli with added noise components using two stimuli, a standard and the odd out pattern (Fine & Jacobs, 2000). In all these cases, judgments potentially related to spatial frequency were improved by training, but the improvements may have involved learning to extract emergent spatial patterns in the stimulus compounds (see also Bennett & Westheimer, 1991). Both cases involved discrimination between two patterns presented in successive temporal intervals within each trial.

Other relevant studies measured learning in spatial-frequency difference thresholds. In one, threshold frequency differences (Δf_t) at $d' = 1.5$ were trained in a yes/no task (standard versus other discrimination using the method of constant stimuli, where the standard was presented on half the trials). Observers showed modest improvements (Meinhardt, 2001), which occurred even when other features of the stimuli were varied (Meinhardt, 2002). Unlike many cases that exhibit long retention, the task needed to be re-learned at 10 months delay (Meinhardt, 2001). Another study showed that training improved performance in both normal and amblyopic observers in a two-interval forced choice spatial frequency discrimination task using standards at two, four, or eight cycles per degree in separate groups (Astle, Webb, & McGraw, 2010). These studies used high contrast stimuli without external noise and showed that spatial frequency judgements could improve over training in binary discrimination tasks.

The current study aims to evaluate perceptual learning in the potentially more challenging spatial frequency judgments with more alternatives and in external noise.

Perceptual learning in *n*-alternative identification

Object identification in the world often requires discriminating between many possibilities. Yet, perceptual learning research has overwhelmingly focused on simpler two-alternative binary choice tasks, with fewer studies on learning in *n*-alternative (n > 2) tasks. In one of these (Gold, Bennett, & Sekuler, 1999; Gold, Sekuler, & Bennett, 2004), observers learned to identify stimuli from a set of either 10 faces or 10 filtered "blob-like" texture patterns embedded in external noise by clicking on the corresponding stimulus in a palette of thumbnail images (see also Hussain, Sekuler, & Bennett, 2009; Hussain, Sekuler, & Bennett, 2011). In another study, observers learned to label multiple artificial creatures (greebles), created from selections of embellishment features, together with names and gender designations (Gauthier, Williams, Tarr, & Tanaka, 1998). In another, improvements in eight-alternative motion direction identification were used to index the effects of exposure to a single "trained" motion direction in the task-irrelevant perceptual learning literature (e.g. Watanabe et al., 2002; Tsushima, Seitz, & Watanabe, 2008). Except for the task-irrelevant learning in motion direction, all these studies almost surely involve multidimensional, not unidimensional, stimulus variations. Here, multidimensional refers to the stimulus variations presented to the observer (e.g. eyes, mouth, and shape variations for faces) and not of the stimulus encodings in the visual system. Absolute unidimensional identification is defined by stimulus variations in one dimension.

In contrast, performance in absolute identification tasks was initially thought to be relatively unaffected by practice (see Dodds et al., 2011 for a review of the prior literature), although sometimes with fewer trials than typically used in perceptual learning studies. The identification literature has predominantly focused instead on performance limits of near three bits of information or seven or fewer categories (the "magical number 7"; Garner & Hake, 1951; Hake & Garner, 1951; Miller, 1956; Shiffrin & Nosofsky, 1994). Still, learning has been reported in a number of cases: identification of line lengths (Rouder, et al. 2004), for angle of inclination of lines, and for dot separation (Dodds et al., 2011). Stimuli were of high contrast and clear visibility. In some of these studies, the number of alternative stimuli *n* was as high as 30, whereas the maximum information transmitted reliably was about 3 to 3.2 bits, or eight to nine responses. Another study trained identification of nine dot separations for 450 trials while varying the distribution of tested stimuli, reporting stimulus distribution context effects and modest learning (Petrov & Anderson, 2005).

N-alternative identification is also an important form of perceptual learning to investigate for several reasons. First, there are many instances in real world applications in which we identify things from larger numbers of alternatives, such as identifying an image as a particular fruit or animal. Second, it is likely to be more efficient in improving the identification of the *n* items than training identification through training stimulus pairs from the set. Third, *n*-alternative identification is more efficient for measuring performance and so for training responses. The guessing rate is quite low compared to two-alternative tasks, so each response (and so each response feedback) offers more learning signal.

Recently, we demonstrated perceptual learning in an absolute orientation identification task tested in visual periphery with external noise and showed how learning depended on different forms of feedback (Liu, Lu, & Dosher, *submitted*). The accuracy of orientation identification improved substantially over training sessions with full response feedback (providing the correct response), some learning occurred with weaker accuracy feedback (indicating only whether the response is correct or not), and a small amount of learning occurred even in the absence of feedback. Learning in these eight-alternative absolute orientation judgments in all three feedback conditions was well accounted for by the identification – integrated reweighting model (I-IRT; Liu, Lu, & Dosher, *submitted*). The I-IRT model in analogous simulations also predicts learning in spatial frequency judgments in similar conditions.

The identification - Integrated reweighting theory

To account for perceptual learning in absolute identification, we extended the original IRT model of perceptual learning that had been developed for binary discrimination (Dosher et al., 2013). The original IRT model accounted for many perceptual learning phenomena, including learning, aspects of transfer, feedback-induced bias, and some failures to learn when stimuli are varied (roved) (see Dosher & Lu, 2017, for a review). Inspired by the neurophysiology, the representation module (signal-processing front end) computes gain-control normalized activity in a set of spatial frequency and orientation sensitive units—the sensory evidence (Petrov, Dosher & Lu, 2005; Petrov, Dosher, & Lu, 2006; Dosher et al., 2013). This representation front end accounts for the effects of contrast and external noise stimulus manipulations on performance.

In the new I-IRT, the observer learns which sensory representation activations to weight toward each identification response. To do this, the model uses a set of mini-decision units, one for each potential response (eight in the current experiments). Figure 2 schematically illustrates this for spatial frequency. As in the earlier IRT, the visual processing "front end" encodes the stimulus as activations in spatial-frequency and orientation-tuned representations at a location-specific level and a location-invariant level that pools over spatial locations. The location-specific representations in the IRT account for specificity of some learning to spatial locations and the location-invariant representations mediate transfer or generalization over spatial testing locations and interactions of learning in multiple locations, both of which occur in perceptual learning tasks (see Dosher et al., 2013; Dosher, Liu, Chu, & Lu, 2020, for reviews). The activations in all these stimulus representations are then weighted to the *n* mini-decision units (one mini-decision unit per response), with the most active mini-decision unit on each trial determining the simulated response (max rule).

After the identification response and any feedback, the weights connecting the sensory evidence (activations in the representation units) to each of the *n* mini-decision units are updated (reweighted) using an augmented Hebbian learning rule. Response feedback (the correct response, the form of feedback used in the current study) shifts activation in each mini-decision unit toward the correct answer for that mini-unit, either match or mismatch, before the weight update. Learning over trials tends to up-weight evidence from its most relevant stimulus representations to each mini-decision unit and down-weight evidence from noisy or irrelevant representations. Performance and learning in the model are controlled by parameters for internal noises, scaling, and nonlinearity, and a single learning rate parameter. The details of the implementation of the I-IRT are briefly described in Appendix A. Using similar parameter values as for orientation identification, the model predicts that learning spatial frequency identification is not only possible but likely.

Perceptual learning in a range of tasks has been modeled using the IRT framework, accounting for many phenomena in perceptual learning in two-alternative tasks (see Dosher & Lu, 2017; Dosher & Lu, 2020 for reviews). In each case, response selection in the model uses the appropriate signal detection analysis for the task. In multi-category identification, a single stimulus is presented, and the max rule is often used, selecting the strongest activation or match. Comparisons of performance across tasks is naturally carried out in the signal detection domain; see Lu and Dosher (2013) section 8.4.

It should also be noted that several cognitive models of absolute identification have been developed previously to account for other aspects of performance that the I-IRT is not designed to account for—such as assimilative or contrastive sequential dependencies

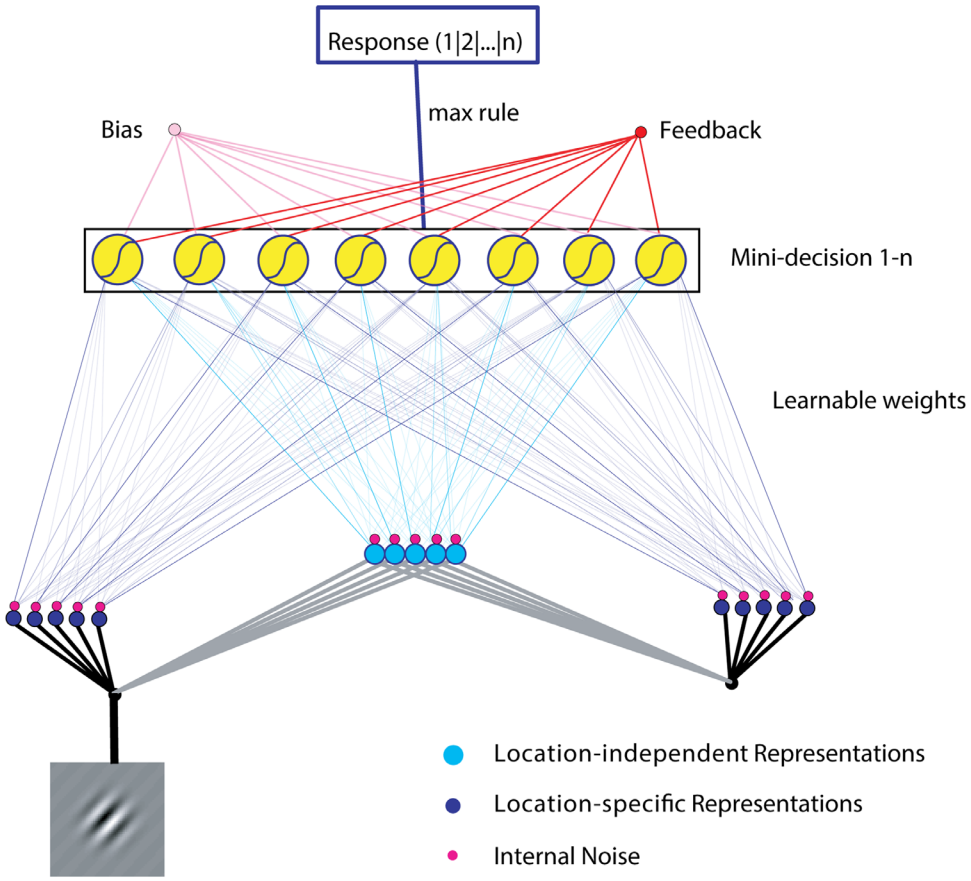

Figure 2. The I-IRT model. A stimulus is first processed into both location-specific and location-independent representations, which are fed forward to *n* mini-decision units. A max rule decides the actual response. Feedback (information about which response is desired) and bias (to balance response frequency for each response) are fed into the mini-decision units, which drive the learning of the model. Learning is achieved through updating weights between representations and mini-decision units.

between trials (Luce, Nosofsky, Green, & Smith, 1982), end-anchor effects (e.g. Ward & Lockhead, 1970; Luce et al., 1982), and response times (see Petrov & Anderson, 2005; Brown, Marley, Donkin, & Healthcote, 2008 for excellent examples). The I-IRT focuses on learning, and it accounts for effects of visual variables in the stimuli, such as contrast, external noise, and stimulus similarity that these earlier cognitive models do not address (the I-IRT does account for some end-anchor effects, see below). Some comparisons to these other models are considered in the Discussion.

Current project

Simulations of the I-IRT predict that learning in spatial frequency identification could occur in many conditions. This hypothesis is tested in two experiments. Experiment 1 examines learning using a mixed-contrast design and measuring accuracy; Experiment 2 examines learning at threshold, in which practice occurs while adaptive measures hold accuracy constant by adjusting stimulus contrast. Both experiments test learning and performance in the presence of modest external noise in the periphery.

Experiment 1

Observers were trained in eight-alternative absolute spatial frequency identification in the periphery at a mixture of target contrasts (0.3, 0.6, and 1.0) using response feedback. Proportion correct is the dependent measure. This design followed the mixed-contrast design used in training absolute orientation identification (Liu, et al., *submitted*). Previous results in perceptual learning suggested that including higher-contrast trials during training can promote learning, at least in two-alternative tasks (e.g., Petrov et al., 2005; Petrov, Dosher, & Lu, 2006; Liu, Lu, & Dosher, 2010; Liu, Lu, & Dosher, 2012). Although some psychophysical tests of detection for spatial frequency stimuli vary the overall stimulus size (number of cycles constant;

e.g. Lesmes et al., 2010), here, we kept the physical size of the Gabor patch the same to minimize the use of stimulus size as a proxy cue for learning, so stimuli for different frequencies show different numbers of spatial cycles.

Methods

Observers

Six observers with normal or corrected-to-normal vision completed the experiment, with written consent under a protocol approved by the Institutional Review Board of the University of California Irvine. Observers participated in eight 960-trial sessions on different days, usually within a 2-week period, for 7680 trials per observer, or 46,080 trials over all the observers. These sessions were run on consecutive days except on weekends, holidays, or occasional scheduling conflicts with observers. They included six sessions of training in external noise and then two sessions in zero external noise to collect data to help constrain the I-IRT model (Observers S1 and S5 performed one or more additional sessions, not analyzed here).

Stimuli and apparatus

A Gabor (windowed sine wave) pattern was presented on each trial at one of the two corners (e.g. top left or bottom right) around fixation with Gaussian noise; its spatial frequency was chosen at random from eight possible values, all shown at orientation θ (see Figure 1 for sample stimuli and an outline of the procedure). The Gabor pattern, defined in a 64 × 64 pixel patch, is described by: $l(x, y) = l_0(1.0 \pm c \sin(2\pi f(y \sin(\theta) \pm x \cos(\theta))) \times \exp(\frac{x^2+y^2}{2\sigma^2}))$, with angle $\theta = 22.5$ degrees (relative to vertical) and one of eight spatial frequencies spaced in half-octave intervals (f = 1/45, 1/32, 1/23, 1/16, 1/11, 1/8, 1/5.7, and 1/4 cycles/pixel, corresponding with 0.43, 0.62, 0.87, 1.23, 1.74, 2.46, 3.48, or 4.92 cycles/degree at the viewing distance), and standard deviation of the Gaussian envelope $\sigma = 0.8$ degrees (16 pixels), maximum contrast c (of 0.3, 0.6, or 1.0), and l_0 is the mid-grey background luminance, with phase as described. Each external noise image, newly generated for each trial and location, was composed of 2 × 2 pixel noise elements with contrasts randomly chosen from a Gaussian distribution with mean value *0* and standard deviation 0.24, and then band-pass filtered (1/16-1/4 cycles/pixel). External noise images and signal Gabor images were displayed sequentially at the frame rate of 60 Hz (see procedure) (NNSSNN). Illustrations of the stimuli (and procedure) are shown in Figure 1.

The images subtended 2.8 degrees × 2.8 degrees visual angle, located at 5.3 degrees eccentricity, at a viewing distance of 83 cm stabilized with a chin rest. Stimuli were generated in MATLAB with PsychToolbox 3 on a Dell PC computer and displayed on a 20-inch Viewsonic color monitor with a refresh rate of 60 Hz and resolution of 640 × 480 pixels in pseudo-monochrome. A lookup table, generated by a psychophysical calibration procedure and validated by photometric measurement, linearized the luminance range into 127 levels from 1 cd/m^2 to 67 cd/m^2; the mid-grey background luminance was 34 cd/m^2.

Design

Observers discriminated the eight spatial frequencies of a Gabor patch in the retinal location (of two locations) indicated by a pre-cue (presented shortly before the Gabor) and a response post-cue. Response feedback (i.e. the correct answer), was provided after each keypress. The Gabor contrast, c, was 0.3, 0.6, or 1.0 (the full contrast range of the display). The number of trials per session (960) was divided equally between the two locations and three contrasts, intermixed randomly within four blocks of 240 trials, yielding 20 tests per stimulus per location per contrast in each session. Because the focus of our experiments and the model was on performance accuracy, instructions emphasized accuracy, and response times were not recorded.

Procedure

Following general instruction on the task, observers were shown examples of the stimuli and performed a small number of practice trials before beginning experimental testing. The goal of the task was to identify the spatial frequency of each image as accurately as possible, and make the best guess when not sure. The instructions explained spatial frequency and that the task was to identify each image as one of eight possible images along a low to high spatial frequency axis. Observers were shown images of the eight spatial frequency stimuli (as in Figure 1) in both zero and high external noise. There were 32 practice trials, two trials per location per spatial frequency, with noise-free stimuli. The practice trials were used to familiarize observers with the experimental set-up, presentation of stimuli, and key presses. Any questions were addressed during the practice.

Each trial started with a central fixation mark and two sets of location markers; 500 ms later the stimulus sequence (external Gaussian noise or blank frames, signal, external Gaussian noise or blank frames) appeared for two refresh counts per frame, with a central pre-cue arrow appearing 100 ms prior to the signal indicating the testing location for that trial. Blank frames replace external noise frames for the two last sessions of testing without external noise. Observers

pressed one of the "a/s/d/f/j/k/l/;" keys, one for each possible spatial frequency, and received response feedback, consisting of a display indicating the correct response, as well as a brief beep if the observer's response was correct. See Figure 1b for an illustration of the trial sequence.

Proportion-correct, psychometric functions, and learning curves

We analyzed the proportion correct data with analysis of variance, as well as on $arcsin\sqrt{p}$ and the logit ($ln(p/(1-p))$, transformations often used to equate the variance and normalize proportions. We also calculated the effect size (η_p^2) and the Bayesian Information Criterion (BIC). The η_p^2 is an effect size measure expressed as proportion variance accounted for after removing the variance of other factors (Bakeman, 2005) and $p_{BIC}(D|H_1)$ is the probability of the BIC given the hypothesis of the effect, H_1, derived from statistics of the analysis of variance (Masson, 2011). In addition, Weibull functions were used to characterize the underlying psychometric functions (proportion correct measured at three contrasts): $\hat{p}_{correct} = p_{max} - p_{min} \times 2^{-(c/\tau)^{\eta}}$, where p_{max} is the upper asymptote of the function, $p_{min} = 1/8 = 0.125$, c is the Gabor contrast, τ is the location (threshold), and η is the slope of the psychometric function. Systems of such functions, one for each session, were fit to the average data, with an assumption of equal slope and equal maximum (1 p_{max}, 1η, 6 τ), using maximum likelihood methods (Wichman & Hill, 2001) and nonlinear minimization routines in Matlab (2019; MathWorks, Natick, MA, USA). The assumption of constant slope is reasonably standard (e.g. Legge, Kersten, & Burgess, 1987; Lu & Dosher, 1999; Hou, Lesmes, Bex, Dorr, & Lu, 2015). The contrast thresholds at a proportion correct of 0.30 were interpolated from the fitted functions for each practice session. Learning curves were then graphed as contrast threshold versus practice session and fit by a power function (Heathcote, Brown, & Mewhort, 2000; Dosher & Lu, 2007): $C(t) = \lambda t^{-\beta} + \alpha$, with initial threshold of $\lambda + \alpha$, asymptotic threshold of α (or a reduced form with $\alpha = 0$), learning rate β, and training block t (least squares methods using nonlinear minimization in Matlab). The proportion of variance accounted for by the power function is r^2:

$$r^2 = 1.0 - \frac{\sum \left[x^{theory} - x^{observed}\right]^2}{\sum \left[x^{observed} - \bar{x}\right]^2}.$$

The Σ is over all N observations and $\bar{x}$ is the mean of the observed values. The learning curves were also tested for significance of learning using F-tests comparing the fits of a fuller model (non-zero learning rate, three parameters) and the nested reduced model (no learning, one parameter equal to the mean): $F(df_1, df_2) = \frac{(r^2_{full} - r^2_{reduced})/df_1}{(1-r^2_{full})/df_2}$, where $df_1 = k_{full} - k_{reduced}$, and $df_2 = N - k_{full} - 1$, where the k's are the number of parameters. The F-test computes the ratio of the improvement in error variance for each additional parameter in the fuller model to the error variance per degree of freedom.

Confusion matrices and weighted-κ

In addition to accuracy and thresholds, an eight by eight confusion matrix was tabulated for each session: $CM(i, j)$ is the frequency of response j to spatial-frequency stimulus i, $i, j \in \{1, 2, \ldots 8\}$. Confusion matrices reveal stimulus confusions, response biases, and improvements across the training process that reduce errors. A Cohen's weighted kappa (κ) was computed from the confusion matrices, which gives full credit to correct responses ("hits"), but also partially credits adjacent responses (a correct response received a weight of 1; the two neighboring responses received a weight of 0.15; and the next more distant two neighboring responses received a weight of 0.05) and corrects for guessing (Cohen, 1968).

Results

Performance accuracy and learning

Figure 3 shows proportion correct as a function of practice session for each of the three contrasts, averaged over observers (panel A). As described in Methods, we performed analysis of variance on $arcsin\sqrt{p}$, a measure which approximately equates the variance and normalizes proportions, to analyze learning over sessions in external noise (sessions 1–6). (Analyses on proportion correct and the logit ($ln(p/(1-p))$ led to essentially equivalent results.) There were significant effects of practice session ($F(5, 25) = 2.830$, $p < 0.0370$, $\eta_p^2 = 0.361$, $p_{BIC}(D|H_1) > 0.999$), contrast ($F(2, 10) = 142.697$, $p < 0.0001$, $\eta_p^2 = 0.966$, $p_{BIC}(D|H_1) > 0.999$), but not of test location or any interactions. Average performance accuracy improved from 31.16 ± 2.22% to 38.72 ± 2.49% across contrast levels and locations, and average accuracies for the three contrast levels were 33.39 ± 1.49%, 39.63 ± 1.56%, and 42.68 ± 2.10%, across sessions and locations.

Learning effects of individual observers, often unreported in the perceptual learning literature, are increasingly being considered. Such analysis can shed light on different learning trajectories and potential causes of failed learning. In this study, perceptual learning was individually significant in four of the six observers (see Appendix B for details). A related study using a somewhat different design found a similar proportion (22 of 30) observers exhibited individually significant learning in spatial frequency identification over five sessions (Liu, Lu, & Dosher, in preparation).

A comparison of the last two high external noise sessions and the two subsequent zero external noise sessions yielded significant effects of external noise ($F(1, 5) = 20.302$, $p < 0.0065$, $\eta_p^2 = 0.8024$, $p_{BIC}(D|H_1) > 0.999$), contrast ($F(2, 10) = 27.517$, $p < 0.0001$, $\eta_p^2 = 0.8462$, $p_{BIC}(D|H_1) > 0.999$), and their interaction ($F(2, 10) = 26.375$, $p < 0.0001$, $\eta_p^2 = 0.8408$, $p_{BIC}(D|H_1) > 0.999$), for $arcsin\sqrt{p}$ (with equivalent results for analyses of proportion correct and the logit transformed proportions). The interaction reflects diminished effects of contrast in zero external noise because of the leftward shift of the psychometric function toward lower contrasts.

Psychometric functions and contrast threshold learning curves

We also estimated the impact of learning on the average psychometric functions (proportion correct as a function of stimulus contrast) by fitting Weibull functions, and a contrast threshold at $p = 0.3$ ($d' = 0.73$) for each session was estimated by interpolation and then fit with a power function (see Methods). Figure 3b shows the Weibull fits to the average proportion correct data from which contrast threshold estimates were derived. Figure 3c shows the threshold data together with a learning curve from the best-fitting reduced power function, with parameters $\gamma = 0.5897 \pm 0.1240$;

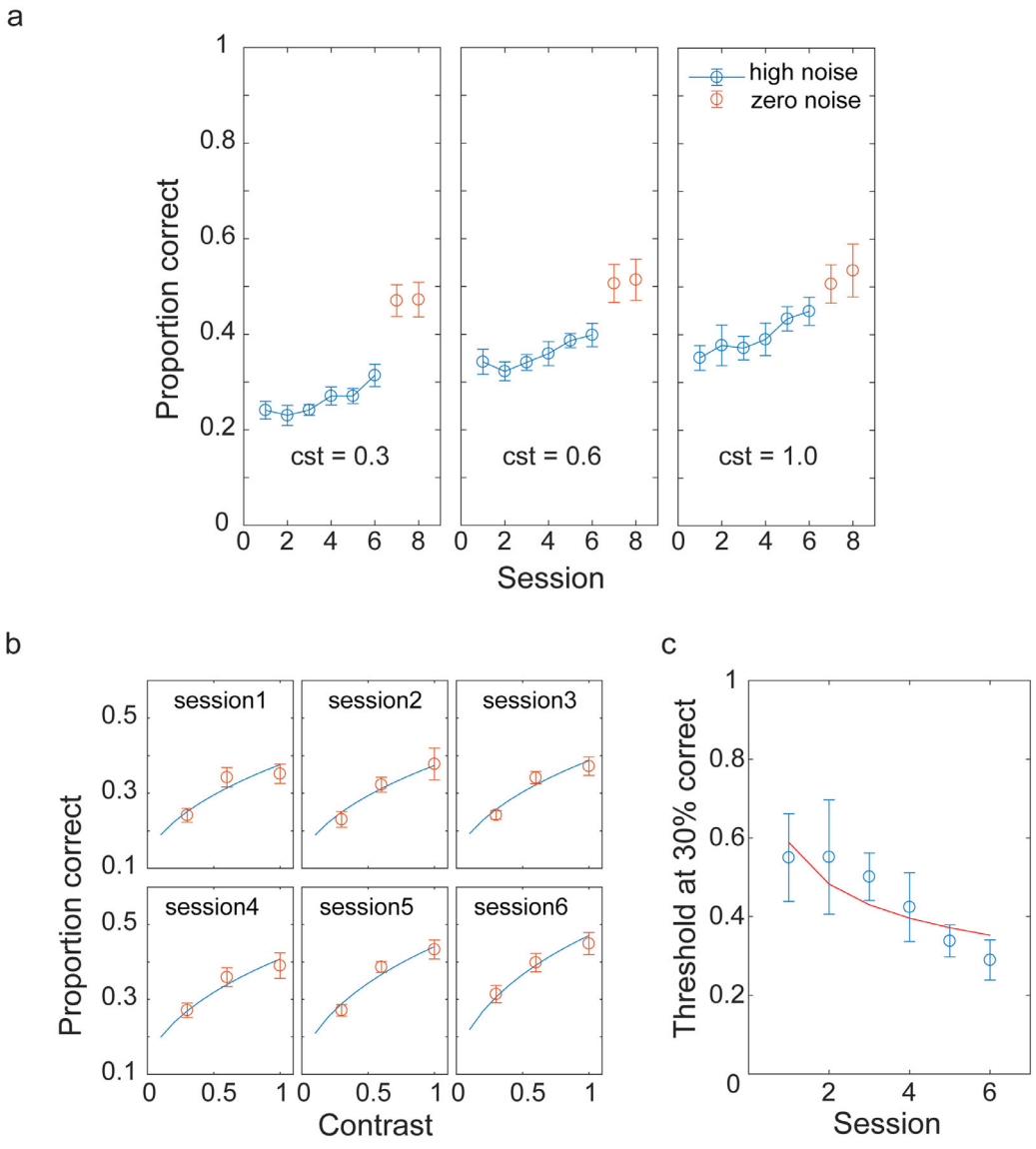

Figure 3. Results of Experiment 1. (**a**) The average proportion correct in three contrast levels over eight sessions, first six sessions in high noise and last two sessions in zero noise for all observers. Observed proportion correct improved for all contrast levels over the sessions. (**b**) The Weibull function fit for average data in each high noise session. (**c**) The 30% correct threshold from the Weibull fit in **b**. The red line is a power function fit of the threshold data. The reduction of thresholds over sessions demonstrates learning.

$\beta = 0.2897 \pm 0.1529$; $\alpha = 0$. Learning curves for data averaged over observers generally take the power form even if the learning curve for individuals is exponential (e.g. Heathcote et al., 2000; Dosher & Lu, 2007; Zhang, Zhao, Dosher, & Lu, 2019a; Zhang, Zhao, Dosher, & Lu, 2019b) because the average curve reflects varied learning parameters. Here, setting $\alpha = 0$ does not reduce the quality of power fit, and when free to vary, the best estimate of α was 0; the value of α (the best contrast threshold achieved after very extensive learning) should be greater than 0, but more data including much longer training would be required to estimate it. The F-test shows the learning fit is marginally better than the no-learning fit ($\beta = 0$): $F(1, 3) = 7.53$, $p = 0.0711$.

Confusion matrices and weighted-κ

The confusion matrix of the eight-alternative task provides complementary information that reveals the similarity structure of the stimuli. Figure 4 shows the aggregate confusion matrices as heatmaps for each training session. Each cell of the heatmap codes the frequency of the response (x-axis, low to high from left to right) for each stimulus (y-axis, low to high from top to bottom). Lighter colors (higher frequencies) on the diagonal show reasonable response tuning in the first session that improves with training. The improvement with training ($f_{i,j}$ (6) − $f_{i,j}$ (1)), right column, shows increased responses on or near the accurate diagonal and decreased responses away from the diagonal. The figure also shows the average confusion matrices from the observers whose performance improved ("learner," S1–S4) and those who did not ("non-learner," S5–S6) in Figure 4b. Compared to the improved responses along the accurate diagonal in learners, non-learners showed more biases – predominantly responding "2" and "7" for most stimuli. Appendix C shows the confusion matrices for each observer.

Correspondingly, the weighted-κ improved from 0.24 to 0.33 over the six training sessions in these average confusion matrices (weighted-κ gives partial credit for near misses and corrects for guessing, with minimum of 0). The confusion matrices in the zero noise sessions on average led to higher weighted-κ (0.43) than in the two previous high external noise sessions, as expected. The statistical details for weighted-κ and for information transmitted (I_t) scores, a common measure of accuracy from the absolute identification literature, are described in Appendix C, and yield similar results as the analysis of the percent correct data. The appendix also considers the slight end-anchor effects, a common observation in the absolute identification literature (see e.g. Ward & Lockhead, 1970; Luce et al., 1982).

Fits of the I-IRT model

We fit the I-IRT model to proportion correct data in the three contrast conditions over training sessions, and the same model and parameters also make predictions about the average confusion matrix data.

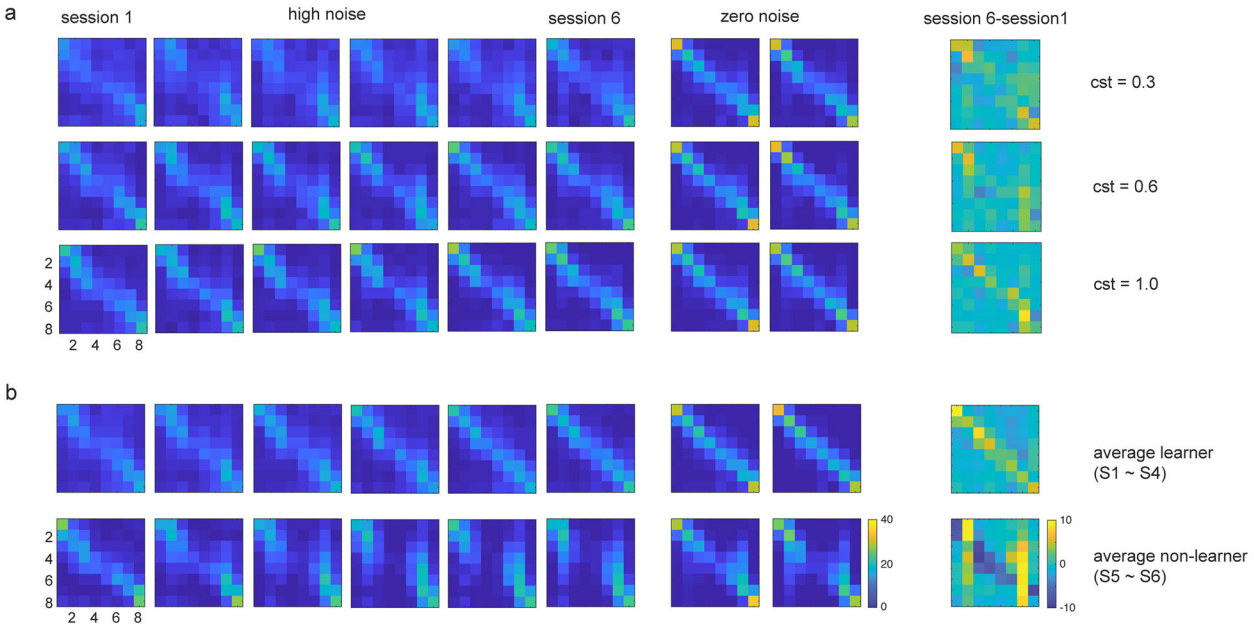

Figure 4. The confusion matrices heatmap in each session, and the change in confusion matrices from session 1 to session 6. The diagonal shows the correct responses. (**a**) Confusion matrices averaged across all observers for each contrast level. Performance was better in higher contrast, in low noise, and in later sessions, seen as cleaner diagonal frequency data. The difference between session 6 and session 1 showed the improved performance. (**b**) Confusion matrices averaged across contrast levels for the average of learners (S1–S4, top), and non-learners (S5–S6, bottom). The difference between session 6 and session 1 showed clear biases in non-learners – who disproportionately responded "2" and "7" for the lower and higher spatial frequencies, respectively.

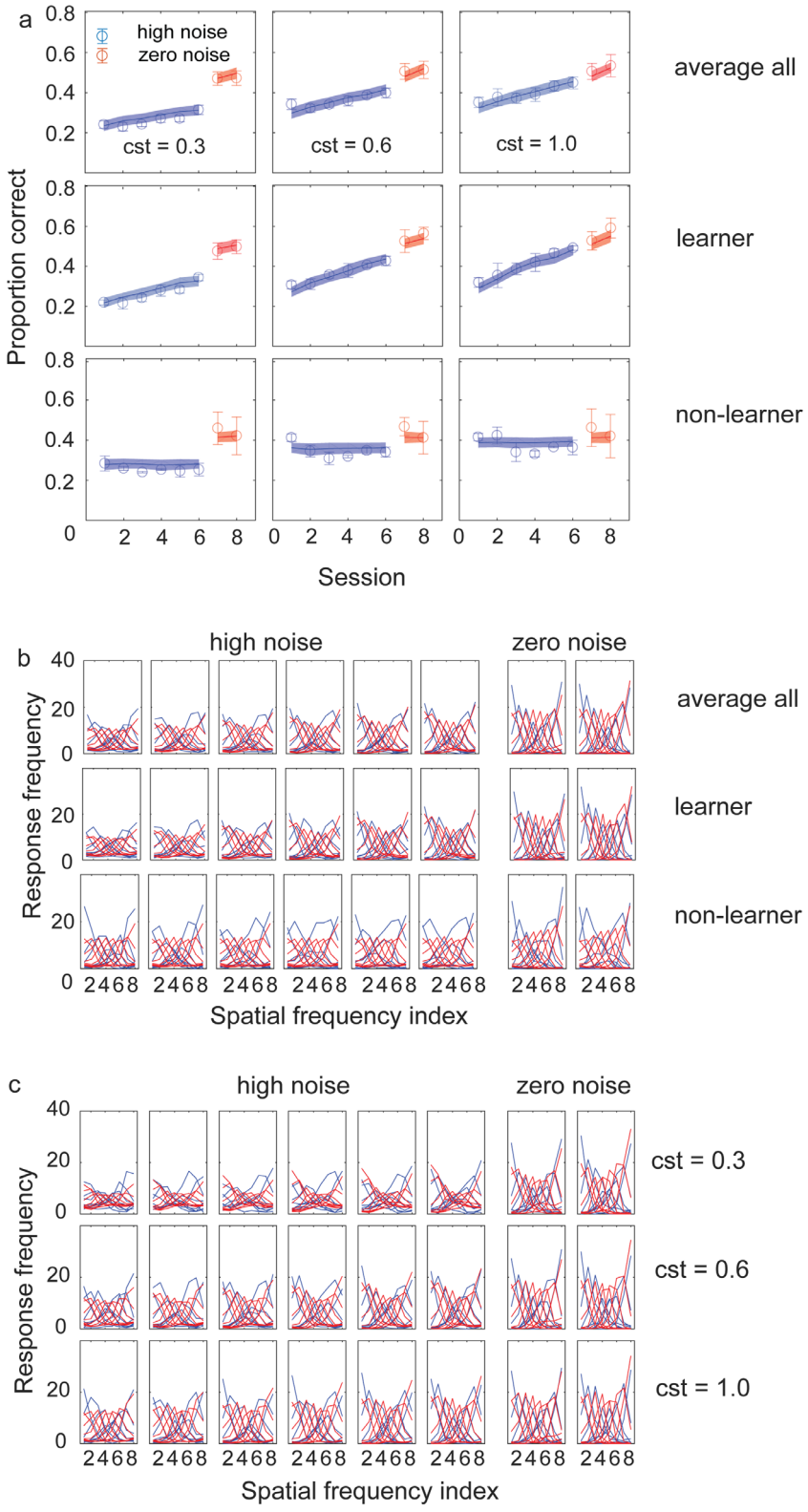

Figure 5. The I-IRT model fit to the experimental data in Experiment 1. (**a**) Model fit to the average of all observers (top), of learners (middle), and of non-learners (bottom). In each case the I-IRT fit the data quite well (see main text for statistics). (**b**) The response function of data (blue) and model predictions (red), averaged over stimulus contrasts, organically emerge from the fit of the model to the proportion correct data in **a**. Each line of the response function shows the frequency of responses to a given stimulus. The model captures the data in non-learners reasonably well even without introducing response biases (see main text for discussion). (**c**) The response function of data (blue) and model (red) for the average of all observers in three contrast levels. The model captures the data in different contrast and external noise levels well.

Parameters		Parameter values	
Parameters set a priori			
Orientation spacing $\Delta\theta$		15 degrees	
Spatial frequency spacing Δf		0.5 octave	
Maximum activation level A_{max}		1	
Weight bounds w_{min} w_{max}		$\pm$1	
Activation function gain γ		3.5	
Location-specific orientation bandwidth h_θ		30 degrees	
Location-independent orientation bandwidth $h_{\theta l}$		60 degrees	
Location-specific frequency bandwidth h_f		1 octave	
Location-independent frequency bandwidth h_{fl}		2 octaves	
Radial kernel width h_r		2 dva	
Parameters adjusted for the data	Average	Learner	Non-learner
Normalization constant k	0	0	0
Scaling factor a	0.1	0.08	0.08
Location-specific internal additive noise, σ_a	5e-7	2e-7	0
Location-independent internal additive noise 2* σ_a	1e-6	4e-7	0
Location-specific internal multiplicative noise, σ_m	0.1	0.1	0.16
Location-independent internal multiplicative noise 2* σ_m	0.2	0.2	0.32
Decision noise σ_d	0.2	0.2	0.1
Learning rate η	1e-4	2e-4	0
Bias weight w_b	0.75	0.75	0.75
Feedback weight w_f	0.75	0.75	0.75
Initial weights scaling factor w_{init}	0.05	0.05	0.05

Table 1. IRT parameters for Experiment 1 (proportion correct).

Figure 5a shows the predictions of the best fitting I-IRT for the average proportion correct data for the three contrast conditions (0.3, 0.6, and 1.0) for the first six training sessions in external noise, and the next two sessions without external noise (line and shaded region at $\pm$ 1σ from the simulations, $n = 100$ simulations; see Appendix A for fitting methods). The model provided an excellent fit ($r^2 = 0.951$) for proportion correct to the effects of learning, contrast, and the interaction with external noise. (Including data both with and without external noise helps constrain internal noise parameters in the model.) The parameters of the fitted model to the average data are listed in Table 1. In addition to fitting the average data across all observers, we also divided the observers into learner and non-learner and fit the average data from both groups (see Appendix B for analyses and discussion of individual observers). The fit of the model to the learner data was excellent ($r^2 = 0.964$ for proportion correct). The fit to non-learner data was also reasonable ($r^2 = 0.822$ for proportion correct) as it still accounts for the effects of contrast, external noise, and their interaction (the lower model r^2 partly reflects the smaller range of data without learning).

Figure 5b shows the confusion matrices, averaged over stimulus contrasts, as response functions – frequencies of responses given a stimulus – for both data and model predictions. Figure 5c shows the response functions for the three contrasts separately for the average of all learners data to illustrate the ability of the model to account for the effects of contrast in the targets and the external noise and the effects of learning, and their interaction (also well accounted for in the learners and non-learners data; $r^2 = 0.820$ for the aggregate data; and $r^2 = 0.843$ for the learner data; $r^2 = 0.536$ for non-learner data).

The predicted response functions of the model were derived from the corresponding model fit to the proportion correct data. That is, the parameters of the model were set to optimize the fit to overall proportion correct, and the predictions for the confusion data emerged organically from that fit. There were no additional parameters for the fit to the confusion matrices.

We offer several comments about the model predictions of confusion matrix data. First, as shown in Figure 5, the model does an excellent job of accounting for the effects of stimulus contrast, external noise, and the improvements over sessions in the response confusion functions. The similarity of the stimuli emerges from the spatial-frequency and orientation-tuned units, and the contrast normalization/gain control in the signal processing front-end.

Second, our observers were generally sensitive to the spatial frequency of the test stimuli in a

task-appropriate way, as seen in the clustering around the correct response diagonal even near the beginning of training (see the heat diagrams for the first session in Figures 4 and 5). This behavioral fact requires partially informative initial starting weights in the model, based on general knowledge of spatial frequency. To implement this, each of the mini-decision nodes was supported by initial positive weight connections from the three most closely matched sets of representation units to the target spatial frequency, with zero initial weights connecting the representation units tuned to other spatial frequencies. Activation in representation units tuned to spatial frequencies above the highest spatial frequency stimulus and below the lowest spatial frequency stimulus naturally combined to predict some of the "end-anchor" effects seen in these data and in other absolute identification data. Still, visual inspection of the data here shows some underperformance relative to the model predictions for the spatial frequency stimuli in the middle of the set and some overproduction of responses at the end points relative to the predicted values, especially in the zero external noise sessions. This may partially reflect response biases especially in non-learners. We explored introducing response biases into the model, either by having different initial weights for different mini-decision units, or having higher response preferences for certain responses (e.g. "2" and "7" as in the non-learners), and these improved the fit to the confusion matrices while yielding a similar fit to the proportion correct data. If the behavioral performance shows a bias profile, it is likely specific to an observer.

Last, we also considered potential corrections for the contrast-sensitivity function on either scaled activation or varied internal noises in the representation units. However, these exploratory simulations had little effect on the fit since all our stimuli were within the more visible central range of spatial frequencies. If stimuli in the reduced sensitivity limbs (high or low) of the contrast sensitivity function were tested, it might require the integration of the contrast-sensitivity function into the I-IRT. Although we investigated introducing response biases or spatial frequency contrast sensitivity corrections in exploratory fits and simulations, we report the fits of the simpler I-IRT.

Figure 6 shows the initial and final (after training) weights from the location-specific and location-invariant representations to each of the eight mini-decision units for fits to all observers and to learners only. Weights of non-learners do not change. As described earlier, initial weights accounted for initial above chance behavioral performance. During learning, reweighting increased the weights on representations most closely tuned to the stimulus represented by that mini-decision unit and reduced the weights on others. The magnitude of the weight changes was higher for learner's data than for the fit to all observers. For mini-decision units representing stimuli in the mid-range of the stimuli, weights increased for units most closely tuned to the corresponding spatial frequency stimulus, and shifted negative for units for nearer competitors—in a classic excitatory center, inhibitory surround pattern. Those units tuned to spatial frequencies yet farther away tended to be slightly negative because the external (white) noise adds distracting activation in all representation units. The mini-decision units for the lowest (or highest) spatial frequency stimuli positively weight units tuned below (or above) which produces the end effects in the predicted response confusion functions. Decision weights on the location-specific and location-invariant representation units were relatively similar although those units are tuned more narrowly and more broadly, respectively (1 and 2 octaves, for full width at half height). Learned weights on location-invariant representations impact performance on (and are trained by) trials in both locations and would be the basis of transfer to untrained locations.

The weights for the fit to non-learners remained unchanged (model learning rate of zero). An alternative fit to non-learners that ignored feedback (feedback weight set to 0) yielded small effects of unsupervised learning and correspondingly small changes in weights. (See the General Discussion for more details.)

Discussion

The I-IRT model predicted that, with response feedback, learning could occur in absolute spatial-frequency identification task with eight alternatives. Behaviorally, we found relatively robust perceptual learning in spatial-frequency identification task in the average data in the intermixed-contrast training paradigm of this experiment. Response confusion data showed substantial sensitivity to the spatial frequency of the stimuli even in the first session that was then further fine-tuned with subsequent training. Robust learning occurred for four of the six observers (each statistically significant individually), whereas the remaining two observers instead developed biased shortcuts counter to the task demands, likely for reasons outside the model framework (see Appendix B for a discussion of individual differences). Even so, the confusion matrices of the non-learners indicate that they were performing some (possibly biased) approximation of the task, as responses continued to carry information about the spatial frequency of the stimulus.

The I-IRT model provided an excellent account of the effects of training, stimulus contrast, and interaction of contrast with external noise on the proportion correct data. It also provided a good account of the confusion data without added model parameters, including the effects of contrast, external noise, learning, and

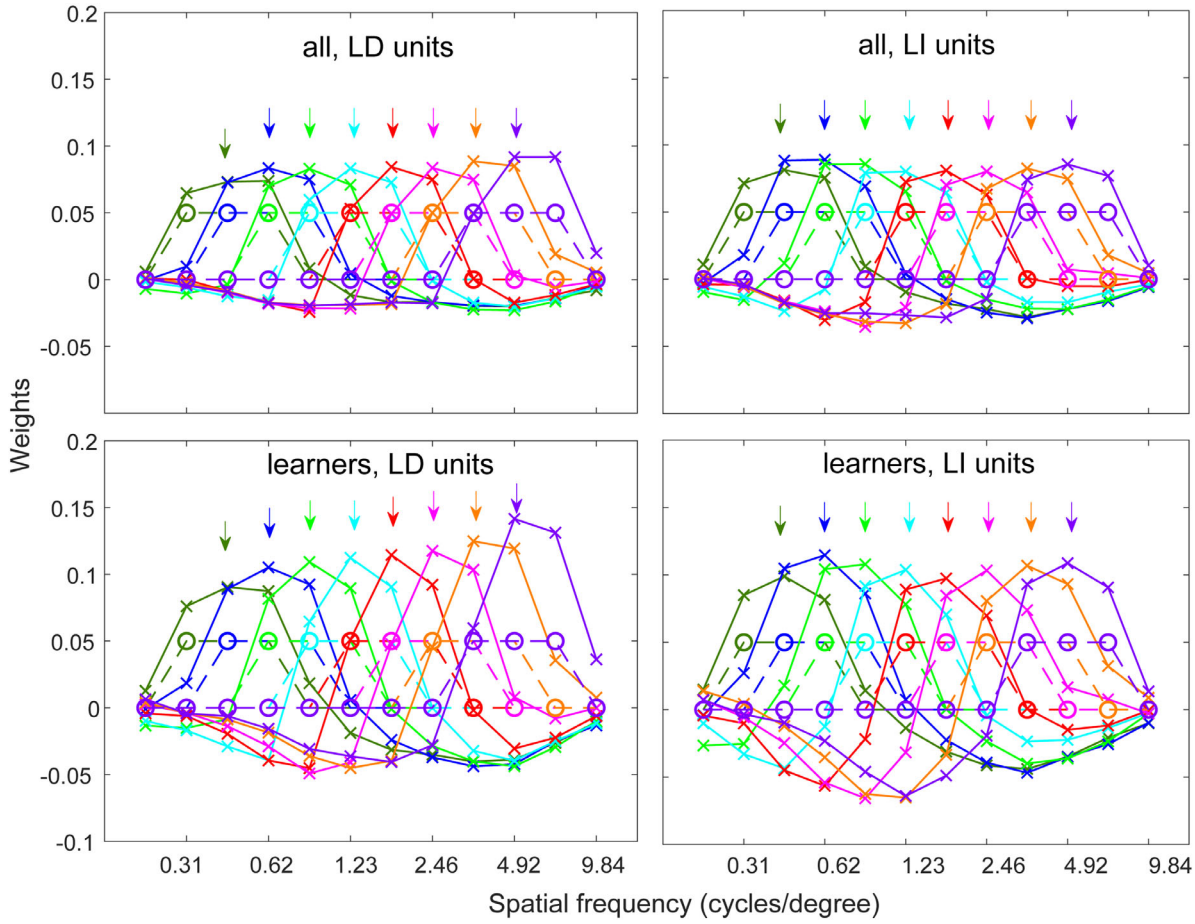

Figure 6. The initial (circles) and final (crosses) weights from representation units to each mini-decision unit for location-specific and location-invariant units for all observers (top panels), and for learners (bottom panels), tuned to the trained orientation. Each color represents one mini-decision unit, and arrows are eight spatial frequencies in the stimuli. Initial weights are set positively for representation units tuned for the spatial frequencies near the stimulus for each mini-decision unit and set to zero for others, corresponding with initial above-chance performance. After learning, the weights of the relevant representation units increased, whereas the weights on other units decreased. For middle spatial frequency stimuli, the pattern shows an excitatory center, inhibitory surround patter to suppress the input from response competitors; for end point stimuli, evidence from representations tuned just outside the range retain positive weights. The weight changes in fits to learners are more robust than for the average of all observers. The weights do not change from initial values in fits to non-learners data. (See text for explanation.)

the degree of confusion between adjacent stimuli. Exploratory fits adding response biases improved the fits to the confusion data slightly but did not alter any conclusions. The learned improvements in performance were the result of dynamic changes in weights on evidence from sensory representations tuned for each mini-decision unit, corresponding with the absolute identification response choices. The I-IRT model can only suggest possible mechanisms for why learning is more robust in some observers (see the General Discussion for further discussion).

Experiment 2

Experiment 2 evaluated learning in spatial frequency identification while training at threshold by adaptively modifying Gabor contrast to track an accuracy of 54%. Practice included six sessions of absolute identification of spatial frequency stimuli in external noise, and an additional two sessions without external noise.

Methods

Observers

Seven observers with normal or corrected-to-normal vision completed the experiment, with written consent under a protocol approved by the Institutional Review Board of the University of California Irvine. Observers participated in 960 experimental trials per session for eight sessions, except for one observer, who completed

only six sessions. Sessions occurred on different days, usually within a 2-week period, for 7680 trials (or 5760 trials) per observer, yielding 51,840 trials over all observers. These sessions were run on consecutive days except on weekends, holidays, or occasional scheduling conflicts with observers. One other observer withdrew after two sessions for scheduling reasons (data not reported).

Stimuli, design, and procedure

The stimuli, design, and procedure were identical to that of Experiment 1, except that the contrasts of the stimuli were set by the accelerated approximation staircase (Kesten, 1958) to track 54% correct (see below) and contrast threshold was the dependent measure. The contrast of the stimulus on the next trial in a test location depended on the accuracy of the prior response in that location through the adaptive staircase. Each 960-trial session was divided into four blocks of 240 trials (120 per location), between which observers could take a short break; new adaptive staircases were restarted using the last contrast tested in the prior block. Observers were encouraged to focus on accuracy and response times were not collected.

Adaptive threshold measurement

The Gabor (signal) contrast on each trial was selected to track a target performance ϕ of 54% correct ($\sim d' = 1.47$ for eight alternatives, standard conversion) using the accelerated stochastic approximation algorithm (Kesten, 1958). In the first two trials, contrasts were determined by the stochastic approximation procedure (Robbins & Monro, 1951): $X_{n+1} = X_n - \frac{s}{n}(Z_n - \phi)$, where n is the trial number, X_n is the stimulus contrast in trial n, $Z_n = 0$ or 1 is the response accuracy in trial n, X_{n+1} is the contrast for the next trial, and s is the pre-chosen initial step size. From the third trial on, the sequence is "accelerated": $X_{n+1} = X_n - \frac{s}{2+m_{shift}}(Z_n - \emptyset)$, where m_{shift} is the number of shifts in response category (from correct response to incorrect response and vice versa). See also Treutwein (1995) and Lu and Dosher (2013) for discussions of the algorithm. We selected $\phi = 54\%$ correct based on pilot data and some simulations. In retrospect a lower value, such as 47% correct ($\sim d' = 1.25$), might have avoided some ceiling effects. The results are unlikely to be changed, however, since reducing the target accuracy reduces contrast and so the stimulus evidence during training. Simulations using target accuracies of both 54% and 47% led to comparable results, whereas a simulation reducing target accuracy to 30% indicated less learning.

Results

Learning functions

The average contrast-threshold learning curve is shown in Figure 7. Contrast thresholds in the first six sessions showed learning in the presence of external noise, whereas the last two sessions were tested with no external noise. An analysis of variance was performed on the thresholds of the first six sessions, with training session and location as experimental factors and observers as the random factor. The main effect of training session was significant ($F(5, 30) = 3.126$, $p < 0.025$, $\eta_p^2 = 0.3425$, $p_{BIC}(D|H_1) > 0.999$), as was

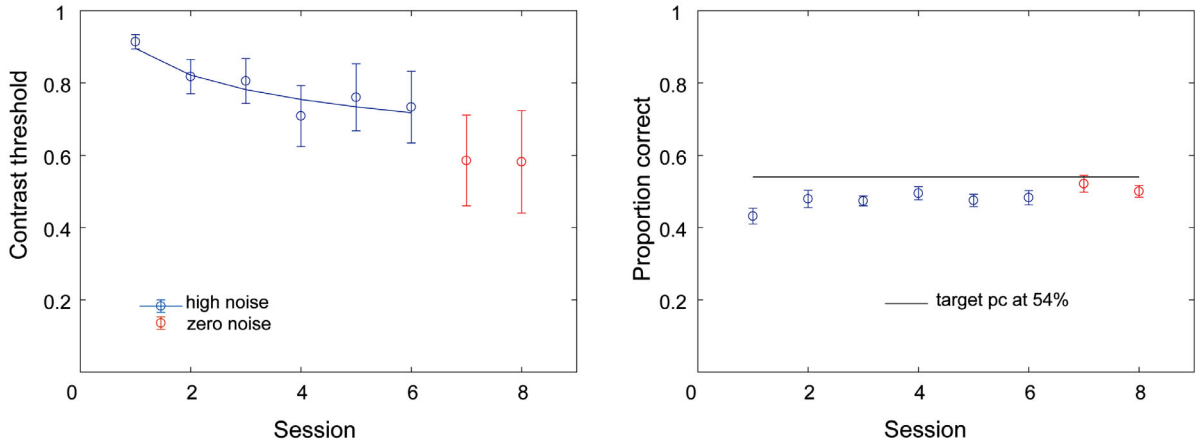

Figure 7. Results for Experiment 2 (contrast threshold version). Left: Contrast thresholds decreased across sessions, showing perceptual learning. Right: Proportion correct over the sessions approximated the 54% target accuracy of the adaptive staircase, although it was somewhat below especially in the first sessions due to ceiling effects in some observers. (See Appendix B for individual observer data.)

location, ($F(1, 6) = 6.259$, $p < 0.05$, $\eta_p^2 = 0.5161$, $p_{BIC}(D|H_1) > 0.999$). The contrast thresholds decreased from 0.914 ± 0.020 to 0.733 ± 0.099 over the six training sessions, showing learning. The performance in the two locations should be similar – the average thresholds across sessions in two locations are 0.764 ± 0.069 and 0.816 ± 0.057, slightly favoring the upper left location for unknown reasons. The best-fitting power function learning curve for the initial training data (sessions 1–6, in external noise) for the average of the observers with an $r^2 = 0.8625$ (initial threshold, $\lambda + \alpha = 0.9063 \pm 0.0252$, $\beta = 0.1280 \pm 0.0707$, α set $= 0$, $p < 0.022$) is shown as the smooth curve. (Setting $\alpha = 0$ did not reduce the quality of the fit in a nested model test; as discussed previously, α should be above 0 but more data from extended training would be necessary to estimate it.) The two sessions in the absence of external noise led to lower thresholds (better performance) than the last two high external noise sessions: 0.584 ± 0.141 vs. 0.747 ± 0.094 ($F(1, 5) = 12.802$, $p < 0.016$, $\eta_p^2 = 0.7191$, $p_{BIC}(D|H_1) > 0.999$).

As in Experiment 1, there was individual variation in perceptual learning. Some observers showed clear learning (S1-S3), or near-ceiling learning (S4, with near-ceiling contrast thresholds with significant improvements in percent correct and in κ), whereas others showed little systematic learning (S5–S7). Learning curves from individual observers are discussed in Appendix B.

Confusion matrices and weighted-κ

Figure 8 shows the average confusion matrices over sessions as a series of heat diagrams. Although the overall accuracy was held approximately constant by the adaptive staircase, and learning was primarily expressed in lowered contrast thresholds, the heat diagrams still showed some improvements in responses near the correct-response diagonal. Weighted-κ, which is calculated from confusion matrices and gives partial credit for close guesses, also increased slightly over the six training sessions in external noise, with means of 0.396, 0.452, 0.443, 0.468, 0.444, and 0.453, (typical standard error $\pm$ 0.02) ($F(5, 30) = 3.328$, $p \approx 0.017$, $\eta_p^2 = 0.3568$, but $p_{BIC}(D|H_1) = 0.481$, with subjects as the random factor). As in Experiment 1, there were end anchor effects, with the accuracy of responses highest (brightest) at the end points of the stimuli. See Appendix C for the corresponding I_t information-transmitted scores and individual differences in the confusion matrices.

Fits of the I-IRT model

We fit the I-IRT model to contrast thresholds over training sessions, which also generated predictions about the response confusion data. The parameters of the best-fitting model are listed in Table 2. In addition to fitting to the average data across all observers, we also performed fits to three classes of observers: learner, ceiling learner, and non-learner. I-IRT fits to data for these three sets of learners are shown in Figure 9. Learning is the natural prediction of the model. The best fitting model to the average data led to an $r^2 = 0.872$ to the thresholds, and $r^2 = 0.751$ for the confusion matrix. The best fitting model to the learner threshold data led to an $r^2 = 0.946$. The same parameters predicted the weighted-κ data with

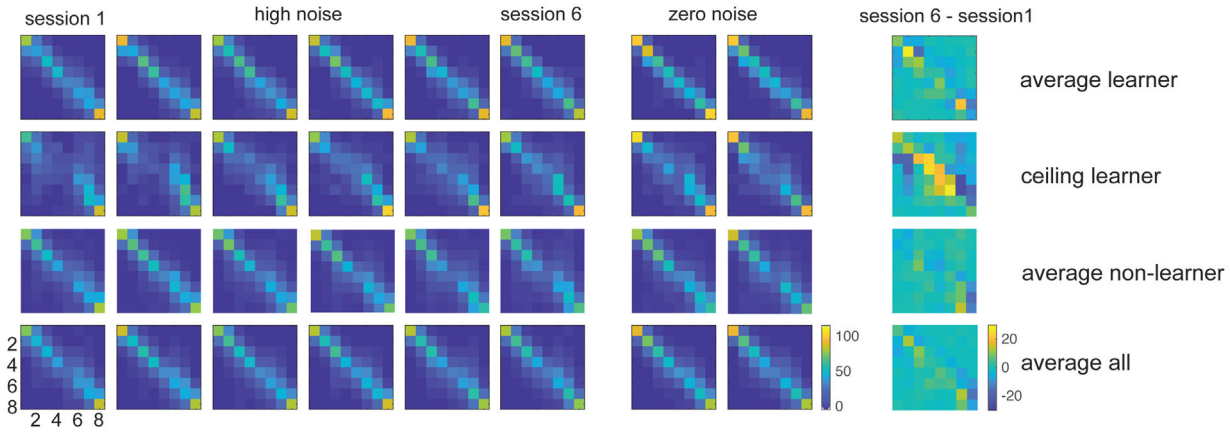

Figure 8. Confusion matrices for Experiment 2 across sessions shown as heatmaps. Confusion matrices are expected to be roughly the same across sessions since the adaptive estimates of contrast threshold sought to hold proportion correct performance at 54%. The data show some improvement over sessions in the confusion matrices because of one near-ceiling observer (see text). (See Appendix B for individual observer data.)

Parameters	Parameter values			
Parameters set a priori				
Orientation spacing $\Delta\theta$	15°			
Spatial frequency spacing Δf	0.5 octave			
Maximum activation level A_{max}	1			
Weight bounds w_{min} w_{max}	$\pm$1			
Activation function gain γ	3.5			
Location-specific orientation bandwidth h_θ	30 degrees			
Location-independent orientation bandwidth $h_{\theta l}$	60 degrees			
Location-specific frequency bandwidth h_f	1 octave			
Location-independent frequency bandwidth h_{fl}	2 octaves			
Radial kernel width h_r	2 dva			
Parameters adjusted for the data	Average	Learner	Ceiling learner	Non-learner
Normalization constant k	0	0	0	0
Scaling factor a	0.1	0.1	0.05	0.1
Location-specific internal additive noise, σ_a	1.5e-5	2e-6	1e-6	5e-7
Location-invariant internal additive noise 2*σ_a	3e-5	4e-6	2e-6	1e-6
Location-specific internal multiplicative noise, σ_m	0.1	0.1	0.1	0.1
Location-invariant internal multiplicative noise 2*σ_m	0.2	0.2	0.2	0.2
Decision noise σ_d	0.09	0.12	0.12	0.08
Learning rate η	3e-5	1e-4	1e-4	0
Bias weight w_b	0.75	0.75	0.75	0.75
Feedback weight w_f	0.75	0.75	0.75	0.75
Initial weights scaling factor w_{init}	0.05	0.05	0.05	0.05

Table 2. IRT parameters for Experiment 2 (threshold) for average, learner, ceiling learner, and non-learners.

$r^2 = 0.723$, and the confusion matrix data with $r^2 = 0.843$. The question is how parameters differ in the ceiling learner and non-learner to account for those data sets. One fit to the ceiling learner threshold data tracked a change in proportion correct ($r^2 = 0.885$), which improved while contrast thresholds were close to ceiling (1.0) across sessions; the same model fit predicted the weighted-κ data, with $r^2 = 0.896$, and predicted the response confusion data with $r^2 = 0.655$. (Note that we are not suggesting that individuals are non-learners in general, but only in this experiment.)

On the other hand, there are many ways to produce little or no learning in the model—failure to use feedback, high internal noises, poor selection of nonlinearity parameter in the decision units, imprecise use of feedback, or the simplest—setting the learning rate to zero. Simulations of several of these explanations were incompatible with some aspect of the data: setting decision noise very high led to changes in the weights that, while they did not yield threshold improvements, still disturbed the default knowledge incorporated in initial weights, disarranging the response confusion matrix in ways not seen in the data. Setting the feedback weight to zero yielded some small amount of unsupervised learning that seemed inconsistent with the data (but showed nearly as good an r^2 compared to no learning, described next). Setting the learning parameter to zero for non-learners led to $r^2 = 0.491$ for threshold data (which has a small range over session since there is little learning yet does express an effect of external noise that the model accounts for), and $r^2 = 0.789$ for the confusion matrix data. However, these simulations are exploratory rather than conclusory.

Figure 10 shows the changes in weights between representation units and each mini-decision unit corresponding to a response. As reported in Experiment 1, weights of the more relevant tuned representation units for each mini-decision unit increased while those of less relevant or most distracting units decreased. This led to an excitatory center and inhibitory surround pattern for mini-decision units associated with mid-range spatial frequency stimuli, and the positive weights pattern for end point units outside the stimulus range (see discussion in Experiment 1). Weight changes in fits to learners were of higher magnitude than in the fits to all observers which included non-learners. Crucially, weights for the ceiling learner (S4) also improved (like the average change of weights), even though the contrast thresholds stayed around the ceiling (1.0) for most of the training period which can be modeled with a lower scaling factor (as in the current simulation) and/or high internal noise than learners.

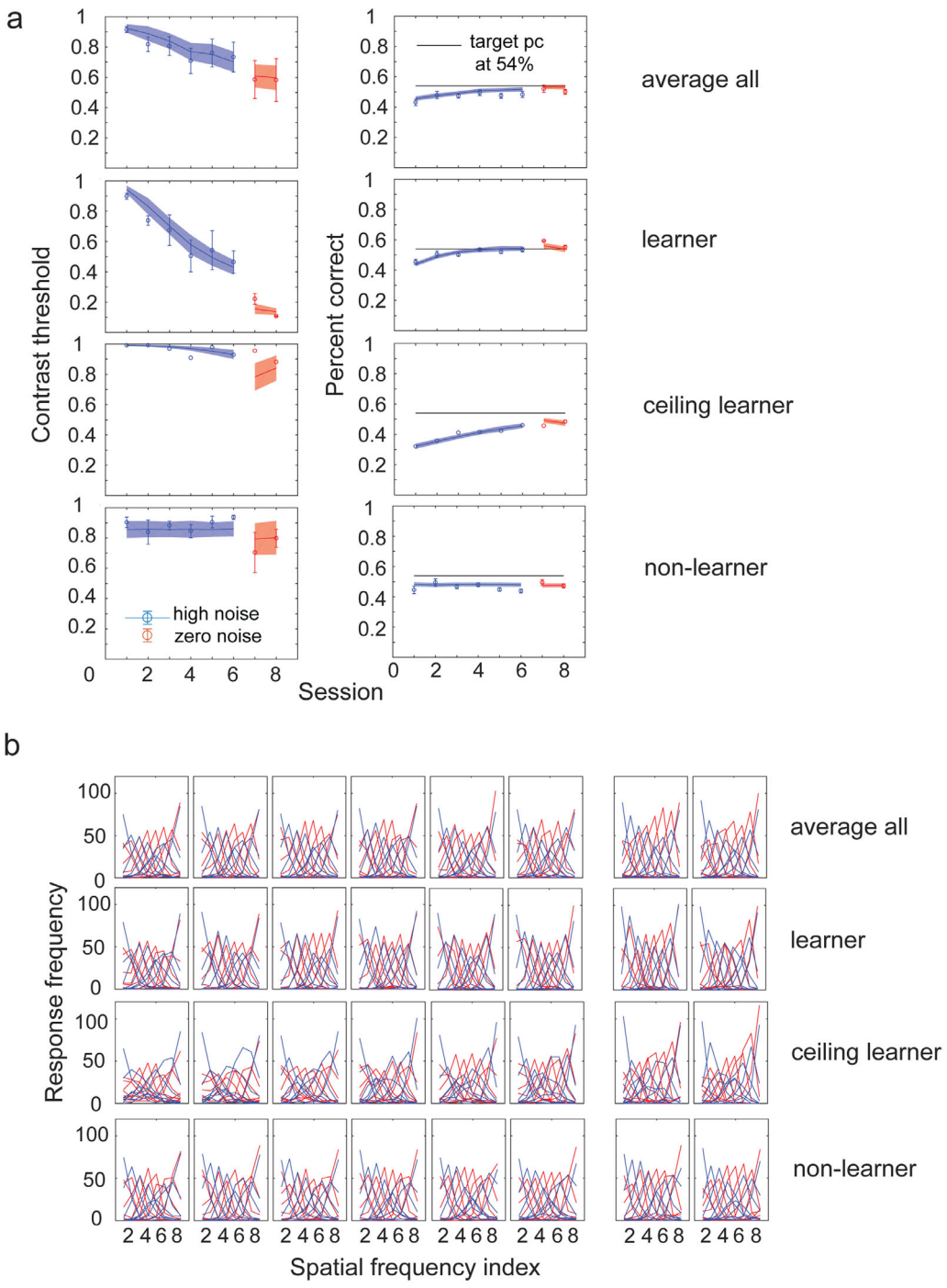

Figure 9. The I-IRT model fit to the experimental data in Experiment 2. (**a**) Model fit to the average across all participants (top), to learners (second row), to the ceiling learner (third row) and to non-learners (bottom). In each case, the I-IRT fit the data quite well (see main text for statistics). (**b**) The response functions of the data (blue) and model (red) based on the fit to threshold data in **a**. Each line of the response function shows the frequency of responses to a given stimulus. The model captures the response function data reasonably well without adding new parameters.

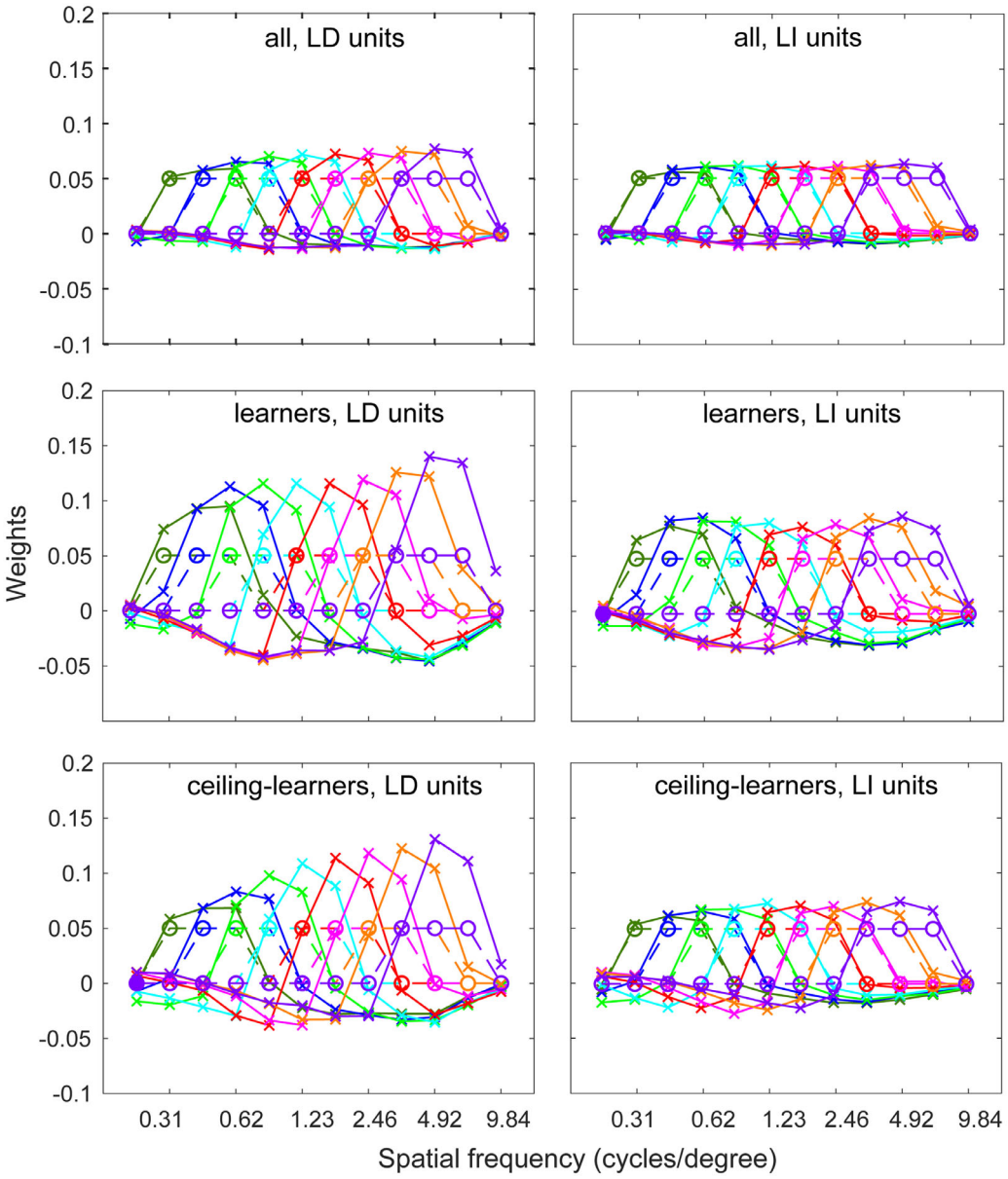

Figure 10. The initial and final weights from representation units to the mini-decision units for location-specific (left) and location-invariant (right) units for the fit to all observers (top panels); for learners only (middle panels); and for the ceiling learner (bottom panels). Learning increased the weights on representations tuned to near-spatial frequencies and decreased the weights on representations tuned to other spatial frequencies, especially those for competing responses. Weights from the fits of the I-IRT simulations changed the most for the learners, slightly less for the ceiling learner, less for the average of all observers, and remained unchanged simulation for non-learners (see text).

Discussion

Perceptual learning occurred in absolute spatial-frequency identification even as learning takes place at near threshold performance levels (ϕ targeting 54% correct). Learning occurred for some, but not all, observers. Three of seven observers showed robust learning, one showed learning at near-ceiling thresholds along with improvements in percent correct and weighted-κ, and three showed either no learning or declines in performance (see Appendix B). All observers, however, demonstrated sensitivity to the stimulus in their confusion matrices (clustering around the correct response diagonal), indicating that they were performing the task, and this was true even at the beginning of training. From this we conclude that perceptual learning in absolute spatial frequency identification is more likely than not to occur in this

paradigm, even when practicing at threshold accuracies and when there are no high-accuracy conditions to improve the rate of learning.

General discussion

Summary

Perceptual learning in spatial frequency tasks has been relatively infrequently studied, and in several cases perceptual learning might have capitalized on emergent spatial pattern features in compound stimulus tasks (see Introduction). In this project, we show that perceptual learning can occur in a challenging absolute spatial frequency identification task with response feedback (full supervision). This is consistent with a related demonstration of perceptual learning in absolute identification of orientation (Liu, Lu, & Dosher, *submitted*). We tested learning of spatial frequency identification in two experiments, one measuring percent correct for stimuli at three contrast levels (including high contrast), and one tracking a threshold proportion correct throughout training by changing the contrast. Robust statistically significant perceptual learning occurred in two-thirds of the observers in the first, and for more than half of the observers in the later. Learning to identify spatial frequency does appear less robust than the learning found in the analogous study of perceptual learning of orientation identification. In that orientation study (Liu, Lu, & Dosher, *submitted*), almost all observers showed substantial learning with response feedback (as used in the current study).

Perceptual learning of spatial frequency

The current study contributes to the modest literature on perceptual learning of spatial frequency judgments, in a task more demanding than yes/no estimates of threshold difference (Meinhardt, 2001; Meinhardt, 2002) or binary identification of compound patterns (Fiorentini & Berardi, 1980; Fiorentini & Berardi, 1981; Fine & Jacobs, 2000). Although there is no perfect test design for investigating spatial frequency discrimination over large ranges, our choice to equate the size of the spatial window of the Gabor patches was necessary to eliminate patch size as a confounded cue during learning. Factors such as differential contrast sensitivity in higher or lower spatial frequencies appeared to have a limited effect in these experiments but might be quite important if stimuli spanned a larger range of spatial frequencies (Campbell, Nachmias, & Jukes, 1970; Campbell & Maffei, 1974).

Perceptual learning of absolute identification

The current study adds another example to the literature demonstrating learning in absolute identification tasks, in which stimuli vary along a single dimension. Previous examples trained features such as line length and line slant or dot distances (e.g, Rouder et al., 2004; Petrov & Anderson, 2005; Dodds et al., 2011) using long stimulus displays. Perceptual learning has been reported in several other *n*-alternative identification tasks involving external noise, such as 10-alternative face/blob identification (Gold, Bennett, & Sekuler, 1999; Gold, Sekuler, & Bennett, 2004), and 10-alternative letter identification (Liu, Lu, & Dosher, 2020). These latter cases, however, surely involve multidimensional stimulus representations, which in principle could provide more opportunities to partitioning the representation space during learning. Several also involve choosing a response by matching to a present template of stimuli, which may serve as another route to a response not typically used in absolute identification tasks.

As discussed above, *n*-alternative identification may be closer to many real-world situations in which visual inputs are identified among a larger number of alternatives. We suggest it is likely more efficient than a training paradigm that reduces the problem to training stimulus pairs from the set. Theoretically, *n*-alternative identification is also more efficient because it offers more learning signal on every trial than two-alternative tasks, due to the much lower guessing rate.

The *n*-alternative absolute identification tasks here trained weights that showed a classical excitatory center, inhibitory surround structure that seeks to optimize identification in the set in a way that pairwise stimulus training would not. Using different pairs to train weights in a series of binary tasks are likely to be slower, or to show the weight-structure disruptions of "roving" paradigms in which the stimuli in binary choice vary (Dosher et al., 2020).

Individual differences in perceptual learning

Individual differences can occur in perceptual learning studies, perhaps more so in certain task domains. The reasons for this include task factors such as task difficulty (Liu et al., 2010; Liu et al., 2012), but may also involve individual differences in underlying abilities as suggested by some researchers (Yang et al., 2020; Dale, Cochrane & Green, 2021).

In our spatial frequency identification experiments, observers were classified as learners or non-learners based on individual-observer significance tests, with consistent converging evidence from response confusion matrices. Our experiments were not designed to probe

the factor(s) that promoted or prevented learning in individuals. Some non-learners developed non-optimal response biases (Experiment 1) while others did not (Experiment 2). All observers carried out the instructed task, in the sense that response confusion data showed some systematic relationship with the stimuli (higher scores along the correct response diagonal), in both learners and non-learners. The distinction between learners and non-learners applies only to the current experiments, and is not intended as a general classification of these individuals. Non-learner individuals here might have learned if trained for a longer time, or under different protocol designs, or in different tasks.

The mix of learners and non-learners in the two experiments reported here was similar to that in other experiments in our laboratory using eight-alternative absolute spatial frequency identification (see Appendix B for details), yielded approximately two-thirds learners. Again, non-learners classified in these studies might not learn in other circumstances, and vice versa.

Identification - Integrated reweighting theory

Learning in the spatial frequency task was predicted by the new I-IRT, initially developed to account for perceptual learning in absolute orientation identification with several forms of feedback (Liu, Lu, & Dosher, *submitted*). The I-IRT as applied to spatial frequency identification is based on the same front-end representation module of orientation- and spatial-frequency tuned representations used to account for perceptual learning in many binary discrimination tasks, as well as the perceptual learning of absolute identification in orientation. To model orientation identification and spatial frequency identification we set up the decision rules for the respective task-relevant judgments. In both cases, learning reweights activation evidence in spatial-frequency and orientation-tuned representations of the input stimuli, making use of feedback supervision during learning (if it is available, as in the current studies). Whether in the three-contrast paradigm, or the threshold learning paradigm, the model provided a good account of the primary indices of learning (increasing proportion correct or decreasing contrast threshold), as well as providing a good account of weighted-κ data and response confusion data with the same parameters. Indeed, the absolute identification paradigm has information advantages relative to simpler binary discrimination tasks in reducing the guessing rate and in providing rich crosschecks on the model, especially the stimulus representations, from confusion matrix data.

In this model, weights connecting to each mini-decision unit start with coarse mapping of spatial frequency information, and these preferences are refined by learning. The use of informed initial weights is consistent with better-than-random initial performance in observers. Starting the model simulations with zero or random weights also yields learning under some training circumstances but predicts close-to-random performance in the beginning and requires quite lengthy training.

Starting from initial weights, learning increases the weights of relevant channels and decreases those of irrelevant channels. The effects of learning were shown in the weight diagrams for the best fitting models. The weights on activation in encoded representation units most closely matched to each spatial frequency response (mini-decision unit) increased; the weights on units supporting adjacent responses (where most errors are made) decreased; and the weights on more distantly tuned units decreased slightly to exclude the external noise response of those units. The initial weight profile incorporated coarser knowledge of spatial frequency corresponding with above chance behavioral performance and structured confusion matrices even at the beginning of training. The weight profiles after training essentially embodied the "template" that has developed for each of the eight response categories.

These weights dynamics are the basis for performance improvements in both low and high external noise, although we primarily examined learning in high external noise in this study. We ran additional simulations to illustrate this point. First, we modeled whether the weight changes from high noise learning would have also benefitted the zero noise stimuli. By freezing initial weights and final weights after training with high external noise, the model estimated contrast thresholds on the same task with zero external noise stimuli "pre-test" and "post-test" and found that the contrast thresholds for zero noise stimuli were reduced after learning with high external noise. So, training in external high noise also benefits the task in zero external noise. Second, we simulated an experiment in which the "pre-" and "post-tests" were done in high external noise, with training with zero external noise, and the model predicts perhaps even more performance improvement in high external noise tests following zero external noise training. Both sets of simulations indicate that learning in the current study is more than just filtering out external noise.

The I-IRT makes predictions about the role of certain task factors in successful learning of absolute identification. For example, it predicts the graded dependence of the amount of learning on the form of feedback: response feedback (as used in the current experiments) is better than accuracy feedback or no feedback (Liu, Lu, & Dosher, *submitted*). The IRT (and the I-IRT) predict the requirement for feedback to enable learning in threshold paradigms tracking lower accuracy levels (Liu et al., 2010; Liu et al., 2012).

The model also predicts the effects on performance of contrast in the signal and in the external noise. Indeed, the IRT framework accounts for many phenomena in perceptual learning.

In the I-IRT, there are several ways for learning to fail in the model. These include ignoring feedback, inability to accurately implement feedback, very high internal noises, inappropriate nonlinearity parameters and response biases, and so on. Each of these mechanisms may yield predicted features in overall performance data or in the confusion data. For example, high internal noise yields overall poor performance and relatively weak behavioral learning, while disarranging the weight structures and confusion data. Choice of atypical nonlinearity parameters in the decision unit may punish performance at different contrast levels differently. Response biases show distinct patterns in the confusion data, and corresponding reductions in performance that mitigate against learning. Imperfect or inappropriate implementation of feedback (in which the information is misapplied to certain mini-decision units) may also result in increased biases and failure to improve performance. Ignoring feedback (by assigning it zero weight) is predicted to potentially yield slight unsupervised learning (depending on initial performance levels, and the presence of high-contrast trials) that could yield some possibly weak learning in an individual observer. We performed several exploratory simulations of these different mechanisms to understand how observers may fail to learn in the model. In the end, there was some feature of the data that eliminated several of these mechanisms, and we elected to model non-learners with a learning rate of zero. A serious investigation of reasons for failure to learn requires further research.

Relation to prior models

Absolute identification has been studied since the 1950s (e.g. Garner, 1953; Miller, 1956), and over that period multiple formal theories have been developed (e.g. Luce et al., 1982; Braida et al., 1984; Petrov & Anderson, 2005; Stewart et al., 2005; and many others). The empirical work (almost always involving clearly visible stimuli and long inspection periods) and the data of these more cognitive theories focused on a variety of behavioral phenomena, including information limits as a function of set size, so-called bow or end-anchor effects on response accuracy, sequential trial effects on errors (either assimilative or contrastive), and the response time of the judgments—with the former two reflecting absolute processing in relation to reference points, and sequential effects reflecting relative processes based on recent trial history. One very nice modeling treatment appears in the SAMBA model (Brown et al., 2008), which integrates aspects of several prior models. It assumes internal distributions of evidence along an abstract representation of the dimension, end anchor effects, relevant decision rules, short-term memory effects, and a ballistic accumulator model of response time and accuracy. That model emphasizes response time data and sequential dependencies and is not focused on modeling learning or the physical stimulus variables of stimulus contrast or external noise. One model (Stewart et al., 2005) assumes that identification is carried out by comparison of the current and prior stimulus and the previous response feedback; a mechanism for learning and of the stimulus contrast and external noise variables would need to be added to explain our effects. Another model (ANCHOR; Petrov & Anderson, 2005) examined some local learning and context effects in addition to these other phenomena and is based on internal codes or anchors for the response categories; this models local learning but not the stimulus variables.

In short, the I-IRT focuses on complementary aspects of performance than these models. It primarily aims to account for learning in absolute identification within a model framework—the integrated reweighting theory (IRT)–that also predicted many of the other phenomena in visual perceptual learning in binary discrimination tasks. That is, the I-IRT has theoretical continuity with the general literature on visual perceptual learning (see Dosher & Lu, 2020, for a review).

There are several points of contrast and comparison to earlier models. (a) The signal-processing front-end of the I-IRT accounts for the effects of stimulus contrast and the effect of external noise, including processes of contrast normalization and internal noise. The nature of the representations also accounts for the response confusion data based on encoding the actual stimulus images into representation activations with intrinsic similarity structure. The assumed spatial-frequency and orientation tuning of the representation units and the (partial optimization of) the weight structure determine the information limits in the task. Within its own domain of visual judgments, the I-IRT could be elaborated to incorporate corrections for the contrast sensitivity function, which might improve the account of the response frequency data across the full visible spatial frequency range. In contrast, the earlier models of absolute identification generally account for complementary phenomena using abstract representations as distributions along the variable dimension and estimate variances of the distributions as parameters (and would require separate set of parameters for each visual condition). (b) The structure of the decision space in the I-IRT accounts for the end-anchor (or bowing) effects of accuracy at the ends of the tested stimulus space and explains why they may not occur in circular dimensions such as orientation. As in those models, the I-IRT could

be extended to incorporate a short-term memory function for recent stimuli to account for potential trial sequential effects. (c) The activation score in each sub-decision unit (corresponding to the possible decisions) could be entered into a competitive system of ballistic accumulators to make the choice, very similar to the implementation for response times in the Brown et al. (2008) model. Any one of these elaborations would involve significant modeling work, as well as new data to constrain them. The I-IRT model as described provided quite a good account of the details of perceptual learning in the current studies.

Conclusions

Significant perceptual learning was demonstrated in a challenging task of absolute identification of spatial frequency with eight alternatives, adding to the small literature on learning in the spatial frequency domain. The existence of learning in an accuracy paradigm and a contrast threshold paradigm, as well as the response frequency distributions (confusion matrices), were both predicted and well-fit by an identification - integrated reweighting theory (I-IRT) simulation model. How the model would predict individual differences between observers remains to be considered. The *n*-alternative identification provides a more efficient paradigm for training expertise and is more representative of some real-world identification tasks.

Keywords: perceptual learning, spatial frequency, absolute identification, learning models

Acknowledgments

Supported by the National Eye Institute (R01EY17491).

Dosher and Liu have no competing interests. Lu holds intellectual property interests in visual function measurement and rehabilitation technologies, and equity interests in Adaptive Sensory Technology, Inc. (San Diego, CA) and Jiangsu Juehua Medical Technology, Ltd (Jiangsu, China); these interests are not related to the current research.

Commercial relationships: none.
Corresponding author: Barbara Dosher.
Email: bdosher@uci.edu.
Address: Department of Cognitive Sciences, University of California, Irvine, 3151 SSPA, Irvine, CA 92697-5100, USA.

References

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature, 387*(6631), 401–406.

Astle, A. T., Webb, B. S., & McGraw, P. V. (2010). Spatial frequency discrimination learning in normal and developmentally impaired human vision. *Vision Research, 50*(23), 2445–2454.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*, 379–384.

Bennett, R. G., & Westheimer, G. (1991). The effect of training on visual alignment discrimination and grating resolution. *Perception & Psychophysics, 49*(6), 541–546.

Braida, L. D., Lim, J. S., Berliner, J. E., Durlach, N. I., Rabinowitz, W. M., & Purks, S. R. (1984). Intensity perception: XIII. Perceptual anchor model of context-coding. *Journal of the Acoustical Society of America, 76*, 722–731.

Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review, 115*(2), 396–425.

Campbell, F. W., & Maffei, L. (1974). Contrast and spatial frequency. *Scientific American, 231*(5), 106–115.

Campbell, F. W., Nachmias, J., & Jukes, J. (1970). Spatial-frequency discrimination in human vision. *Journal of the Optical Society of America, 60*(4), 555–559.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220.

Crist, R. E., Kapadia, M. K., Westheimer, G., & Gilbert, C. D. (1997). Perceptual learning of spatial localization: Specificity for orientation, position, and context. *Journal of Neurophysiology, 78*(6), 2889–2894.

Dale, G., Cochrane, A., & Green, C. S. (2021). Individual difference predictors of learning and generalization in perceptual learning. *Attention Perception & Psychophysics, 83*, 2241–2255.

Dodds, P., Donkin, C., Brown, S. D., & Heathcote, A. (2011). Increasing capacity: Practice effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 477–492.

Dosher, B., 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., Liu, J., Chu, W., & Lu, Z.L. (2020). Roving: The causes of interference and re-enabled learning in multi-task visual training. *Journal of Vision, 20*(6), 9.

Dosher, B., & Lu, Z.-L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences, 95*(23), 13988–13993.

Dosher, B., & Lu, Z.-L. (1999). Mechanisms of perceptual learning. *Vision Research, 39*(19), 3197–3221.

Dosher, B., & Lu, Z.-L. (2007). The functional form of performance improvements in perceptual learning: Learning rates and transfer. *Psychological Science, 18*, 531–539.

Dosher, B., & Lu, Z.-L. (2017). Perceptual learning and models. *Annual Review of Vision Science, 3*, 343–363.

Dosher, B., & Lu, Z.-L. (2020). *Perceptual learning: How experience shapes visual perception*. Cambridge, MA: MIT Press.

Fahle, M. (2005). Perceptual learning: Specificity versus generalization. *Current Opinion in Neurobiology, 15*(2), 154–160.

Fahle, M., Edelman, S., & Poggio, T. (1995). Fast perceptual learning in hyperacuity. *Vision Research, 35*(21), 3003–3013.

Fine, I., & Jacobs, R. A. (2002). Comparing perceptual learning across tasks: A review. *Journal of Vision, 2*(2), 5–5.

Fine, I., & Jacobs, R. A. (2000). Perceptual learning for a pattern discrimination task. *Vision Research, 40*(23), 3209–3230.

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

Fiorentini, A., & Berardi, N. (1981). Learning in grating waveform discrimination: Specificity for orientation and spatial frequency. *Vision Research, 21*, 1149–1158.

Fiorentini, A., & Berardi, N. (1997). Visual perceptual learning: A sign of neural plasticity at early stages of visual processing. *Archives Italiennes de Biologie, 135*(2), 157–167.

Garner, W R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology, 46*, 373–380.

Garner, W. R., & Hake, H. W. (1951). The amount of information in absolute judgments. *Psychological Review, 58*(6), 446.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training 'greeble' experts: A framework for studying expert object recognition processes. *Vision Research, 38*(15–16), 2401–2428.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. *Nature, 402*(6758), 176–178.

Gold, J. M., Sekuler, A. B., & Bennett, P. J. (2004). Characterizing perceptual learning with external noise. *Cognitive Science, 28*(2), 167–207.

Hake, H. W., & Garner, W. R. (1951). The effect of presenting various numbers of discrete steps on scale reading accuracy. *Journal of Experimental Psychology, 42*(5), 358.

Heathcote, A., Brown, S., & Mewhort, D. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*, 185–207.

Hou, F., Lesmes, L., Bex, P., Dorr, M., & Lu, Z.-L. (2015). Using 10AFC to further improve the efficiency of the quick CSF method. *Journal of Vision, 15*(9):2, 1–18.

Hussain, Z., Sekuler, A. B., & Bennett, P. J. (2009). How much practice is needed to produce perceptual learning? *Vision Research, 49*(21), 2624–2634.

Hussain, Z., Sekuler, A. B., & Bennett, P. J. (2011). Superior identification of familiar visual patterns a year after learning. *Psychological Science, 22*(6), 724–730

Karni, A., & Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences, 88*(11), 4966–4970.

Kesten, H. (1958). Accelerated stochastic approximation. *The Annals of Mathematical Statistics, 29*(1), 41–59.

Legge, G. E., Kersten, D., & Burgess, A E. (1987). Contrast discrimination in noise. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 4*, 391–404.

Lesmes, L. A., Lu, Z. L., Baek, J., & Albright, T. D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method. *Journal of Vision, 10*(3), 17–17.

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

Liu, J., Lu, Z.-L., & Dosher, B. (2010). Augmented Hebbian reweighting: Interactions between feedback and training accuracy in perceptual training. *Journal of Vision, 10*(10), 29.

Liu, J., Lu, Z.-L., & Dosher, B. (2012). Mixed training at high and low accuracy levels leads to perceptual learning without feedback. *Vision Research, 61*, 15–24.

Liu, J., Lu, Z. L., & Dosher, B. (2020). Similar perceptual learning in 10-alternative letter identification in external noise with and without feedback supervision. *Journal of Vision, 20*(11), 1237–1237.

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

Liu, Z. (1999). Perceptual learning in motion discrimination that generalized across motion directions. *Proceedings of the National Academy of Sciences, 96*(24), 14085–14087.

Lu, Z., & Dosher, B. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 16*, 764–778.

Lu, Z.-L., & Dosher, B. (2004). Perceptual learning retunes the perceptual template in foveal orientation identification. *Journal of Vision, 4*(1):5, 44–56.

Lu, Z.-L., & Dosher, B. (2013). *Visual psychophysics: From laboratory to theory*. Cambridge, MA: MIT Press.

Lu, Z.-L., & Dosher, B. (2022). Current directions in visual perceptual learning. *Nature Reviews Psychology, 1*, 654–668.

Luce, R. D., Nosofsky, R. M., Green, D. M., & Smith, A F. (1982). The bow and sequential effects in absolute identification. *Perception & Psychophysics, 32*(5), 397–408.

Masson, M E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*, 679–690.

Meinhardt, G. (2001). Learning a grating discrimination task broadens human spatial frequency tuning. *Biological Cybernetics, 84*, 383–400.

Meinhardt, G. (2002). Learning to discriminate simple sinusoidal gratings is task specific. *Psychological Research, 66*, 143–156.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for information processing. *Psychological Review, 63*, 81–97.

Petrov, A. A., & Anderson, J R. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review, 112*(2), 383–416.

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

Petrov, A. A., Dosher, B., & 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*(5059), 1018–1021.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics, 22*, 400–407.

Rouder, J. N., Morey, R. D., Cowan, N., & Pealtz, M. (2004). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin & Review, 11*(5), 938–944.

Saarinen, J., & Levi, D M. (1995). Perceptual learning in Vernier acuity: What is learned? *Vision Research, 35*(4), 519–527.

Sagi, D. (2011). Perceptual learning in vision research. *Vision Research, 51*(13), 1552–1566.

Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review, 101*(2), 357–361.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*, 881–911.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35*, 2503–2522.

Tsushima, Y., Seitz, A. R., & Watanabe, T. (2008). Task-irrelevant learning occurs only when the irrelevant feature is weak. *Current Biology, 18*(12), R516–R517.

Ward, L. M., & Lockhead, G R. (1970). Sequential effects and memory in category judgments. *Journal of Experimental Psychology, 84*(1), 27–34.

Watanabe, T., Náñez, J. E., Sr, Koyama, S., Mukai, I., Liederman, J., & Sasaki, Y. (2002). Greater plasticity in lower-level than higher-level visual motion processing in a passive perceptual learning task. *Nature Neuroscience, 5*(10), 1003–1009.

Watanabe, T., & Sasaki, Y. (2015). Perceptual learning: Toward a comprehensive theory. *Annual Review of Psychology, 66*, 197.

Wichman, F.A., & Hill, N J. (2001). The psychometric function I: Fitting, sampling, and goodness-of-fit. *Perception & Psychophysics, 63*(8), 1293–1313.

Yang, J., Yan, F. F., Chen, L., Xi, J., Fan, S., Zhang, P., . . . Huang, C.-B. (2020). General learning ability in perceptual learning. *Proceedings of the National Academy of Sciences, 117*(32), 19092–19100.

Yu, C., Klein, S. A., & Levi, D. M. (2004). Perceptual learning in contrast discrimination and the (minimal) role of context. *Journal of Vision, 4*(3), 4.

Young, K. G., Li, R. W., Levi, D. M., Klein, S. A., & Huang, E. Y. (2004). Interocular transfer in

perceptual learning of a Vernier task. *Investigative Ophthalmology and Visual Science, 45*(13), 4363–4363.

Zhang, P., Zhao, Y., Dosher, B., & Lu, Z.-L. (2019a), Assessing the detailed time course of perceptual sensitivity change in perceptual learning. *Journal of Vision, 19*(5), 9.

Zhang, P., Zhao, Y., Dosher, B., & Lu, Z.-L. (2019b). Evaluating the performance of the staircase and quick Change Detection methods in measuring perceptual learning. *Journal of Vision, 19*(7), 14.

Zhou, Y., Huang, C., Xu, P., Tao, L., Qui, Z., & Li, X. et al. (2006). Perceptual learning improves contrast sensitivity and visual acuity in adults with anisometropic amblyopia. *Vision Research, 46*(5), 739–750.

Appendix A: The identification - Integrated reweighting theory

The original integrated reweighting theory (IRT) (Dosher et al., 2013) was developed to account for perceptual learning and transfer over changes in stimulus (e.g. changed orientation) or changed spatial location in two-alternative pattern discrimination tasks. The I-IRT model (Liu, Lu, & Dosher, *submitted*) extends predictions to identification by introducing *n* mini-decision units, one for each possible response. The final response on a trial is based on the mini-decision unit with the highest activation. As with the original IRT, the input stimulus is encoded as a pattern of activity in spatial-frequency and orientation tuned representation units at both location-specific and location-independent levels. The model uses hybrid learning rules: unsupervised Hebbian learning augmented by feedback supervision when available (from the augmented Hebbian reweighting model, or AHRM; Petrov et al., 2005; Petrov et al., 2006), which accounts for learning outcomes with and without feedback. The simulations mimic exactly the details of the experiments, using the same program to generate stimulus images, numbers of trials and randomization, etc. and the simulated data are then processed as in the behavioral experiment (here, as proportion correct or threshold, and confusion matrices). The model is implemented in Matlab (The MathWorks, Inc., Torrance, CA, USA). We briefly summarize the model here, including equations and descriptions found in the original IRT papers (Dosher et al., 2013; Liu et al., 2014; Liu, Dosher, & Lu, 2015). The descriptions of model equations below are necessarily similar to treatments in Dosher et al. (2013), and other papers using the two-alternative IRT.

The I-IRT, like the IRT, has a representation module, a decision module, and a learning module. The representation module processes the stimulus images from the experiment to compute the activities in location-specific and location-invariant representations (sometimes cited as analogous to the visual areas V1 and V4 or IT). The input image, $I(x, y)$, is defined as the sum of the signal and noise images for a given trial, corresponding with an integration of noise and signal frames through temporal integration by the visual system. The input image is then convolved with the filter characterizing each spatial-frequency/orientation tuned representation unit using a fast Fourier transform, followed by half-squaring rectification, to produce phase-sensitive activation maps analogous to "simple cells":

$$S(x, y, \theta, f, \phi) = [RF_{\theta,f,\phi}(x, y) \otimes I(x, y)]^2_+.$$

In this implementation the orientation/spatial-frequency filters at each spatial point sample 12 spatial frequency bands (every 1/2 octave) centered at [0.22 0.31 0.43, 0.62, 0.88, 1.23, 1.75, 2.46, 3.51, 4.92, 7.01, and 9.85 cycles/degree] × 12 orientation bands (every 15 degrees) centered at [0°, ±15°, ±30°, ±45°, ±60°, ±75°, and +90° (=−90°)] × four spatial phases [0°, 90°, 180°, and 270°]. In the location-specific representations, the spatial frequency tuning is set at $h_f = 1$ octave and the orientation tuning is set at $h_\theta = 30$ degrees (half-amplitude full-bandwidth), based on cellular physiology in primary visual cortex. In the location-invariant representations, bandwidths were set at twice those of the location-specific units since cells in higher visual areas are more broadly tuned to spatial frequency and orientation; they also may have more internal noise. (These representation module parameter values have been used in many model applications in the IRT framework.) The phase-sensitive maps $S(x, y, \theta, f, \phi)$ are pooled over spatial phases to create phase-invariant energy maps: $E(x, y, \theta, f) = \Sigma S(x, y, \theta, f, \phi) + \varepsilon_1$, where ε_1 is an internal Gaussian noise (mean 0, standard deviation σ_1). These maps include nonlinear inhibitory normalization: $C(x, y, \theta, f) = \frac{aE(x,y,\theta,f)}{k+N(f)}$. The normalization pool $N(f)$ sums over all orientations, with slight tuning for similar spatial frequencies, consistent with physiological and psychophysical evidence. The saturation constant k avoids division by zero at very low contrasts when the normalization pool is very small, and can be set to 0 in the current experiment which uses medium-to-high contrasts. The parameter a is a scaling factor that can shift the range of the final activation values.

To compress the number of representations, the normalized phase-insensitive maps $C(x, y, \theta, f)$ are pooled over space around the target stimulus with a Gaussian kernel of radius W_r, and Gaussian additive noise is added to the system (mean 0 and standard deviation σ_2): $A'(\theta, f) = \sum_{x,y} W_r(x, y) C(x, y, \theta, f) + \varepsilon_2$. Then, a nonlinear function with

a gain parameter γ limits the activations of each representation to the range of (0, A_{max}):

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

Large caches of activation patterns over these representations are computed for different contrasts and samples of external noise for use in the trial-by-trial simulations of the experiments.

The decision module uses eight mini-decision units, one for each spatial-frequency response. On every trial, each mini-decision unit is driven by the weighted activation from the representation units, input from a bias unit, and internal noise, leading to a noisy decision variable: $u_i = \sum_{j=1}^{96} w_{ji} A(\theta_{ji}, f_{ji}) - w_b b_i + \varepsilon_d$. The w_{ji} values are the current weights connecting representation units to sub-decision unit i, b is a bias term weighted by w_b and ε_d (Gaussian, mean 0, standard deviation σ_d) is the (same) decision noise for each sub-decision unit i. A sigmoidal function with parameter γ transforms this into the "early" post-synaptic decision activation: $o_i' = G(u_i) = \frac{1-e^{-\gamma u_i}}{1+e^{-\gamma u_i}} A_{max}$. A maximum rule selects the final response.

The learning module updates the weights between the representation units and the mini-decision units on every trial. The decision variable in each mini-decision unit u_i is shifted towards the correct response (provided by the response feedback in the experiments) to generate the "late" post-synaptic activation: $o_i = G(u_i + w_f F)$, which moves the weights in the right direction. With a high feedback weight w_f, the "late" decision activation approaches the correct output ($\pm A_{max} = \pm 1$), which in turn improves learning. (If no feedback signal is available, which never occurred in the experiments here, $F = 0$, and learning relies on the unsupervised early decision value ($o = o'$), which can often be less efficient.) Feedback was implemented at each of the mini-decision units. Because the feedback provided the correct response, the algorithm always sets $F = 1$ for the mini-decision corresponding to the correct response and $F = -1$ for all the other mini-decision units, regardless of whether the observer's response is correct or not.

Weight changes are determined by: $\Delta w_i = (w_i - w_{min})[\delta_i]_- + (w_{max} - w_i)[\delta_i]_+$, whereas $\delta_i = \eta A(\theta_i, f_i)(o - \bar{o})$, where $A(\theta, f)$ is the pre-synaptic activation, and $(o - \bar{o})$ is the difference between the post-synaptic activation and its long-term average $\bar{o}$ weighted exponentially over the last 50 trials: $\bar{o}(t+1) = \rho\, o(t) + (1-\rho)\bar{o}(t)$, $\rho = 0.02$, w_{min} and w_{max} are the lower and upper bounds of weights (to prevent weights exploding). Each mini-decision unit also receives input from a bias term b to balance the response frequencies, also exponentially time-weighted with a time constant 0.02: $r(t+1) = \rho\, R(t) + (1-\rho)\bar{r}(t)$, $b(t+1) = r(t)$. Here, $R(t)$ is 1 for the actual response, and $-1/7$ for the other potential responses. The bias input works against unbalanced response frequencies.

The I-IRT model was fit to the data, whether proportion correct or threshold, by varying key parameters of the model, simulating the model 100 times, carrying out the same data analysis as for the behavioral data, and then comparing the mean simulated outcomes with the data. Most of the parameters were set a priori from prior applications of the IRT model, originally motivated by the physiology. A grid of parameter values (noise terms, scaling factor, model learning rate, and initial weights) was evaluated, centered around values from previous fitted applications of the model and spanning around a 10 times to 100 times range. Then the parameter space was heuristically searched in a finer grid in the regions yielding higher quality of fit (least squared errors). Occasionally, when multiple parameter combinations yielded satisfactory fits (equivalent r^2), the one most consistent with prior applications was selected. The simulated results are shown in as a region with ± 1 standard deviation of the mean prediction computed from the 100 learning curves simulated from the set of best-fitting parameter values, with the quality of the fit summarized by the r^2.

Appendix B: Individual observer data

Experiment 1

Experiment 1 demonstrated solid perceptual learning in the 8-alternative spatial frequency task using a three contrast (0.3, 0.6, and 1.0) design. As is common in perceptual learning, there was also some evidence for individual variation. Analyses of individual data revealed that four observers (S1–S4) showed robust learning while two (S5–S6) did not (see Figure B1). To assess this for each observer individually, we compared the proportion correct in the first ($\hat{p}_1$) and the last ($\hat{p}_6$) training sessions as a measure of learning for each contrast condition (pooled over location) ($z = (\hat{p}_6 - \hat{p}_1)/\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_6})}$, where $\hat{p}$ is the average proportion correct, $n_1 = n_6 = 320$). Consistent with the visual impression, learning was significant for four of the six observers, for all contrasts (0.3, 0.6, and 1.0, respectively) (S1: $z = 3.032$, $p < 0.0012$; $z = 3.143$, $p < 0.0008$; $z = 5.8327$, $p < 0.00001$; S2: $z = 3.360$, $p < 0.0004$; $z = 4.693$, $p < 0.0001$; $z = 3.515$, $p < 0.00002$; S3: $z = 3.033$, $p < 0.001$; $z = 1.302$, $p < 0.096$; $z = 4.045$, $p < 0.00001$; S4: $z = 4.636$, $p < 0.0001$; $z = 3.508$, $p < 0.0002$; $z = 4.394$, $p < 0.00001$; S5: $z = -1.032$, $p \approx 0.849$; $z = -1.696$, $p \approx 0.955$; $z = -0.882$, $p \approx .0.811$; S6: $z = -0.747$, $p > 0.7$; $z = -2.066$, $p > 0.9$; $z = -1.809$, $p > 0.9$). (Consistent results were found using analysis of variance for individual observers using

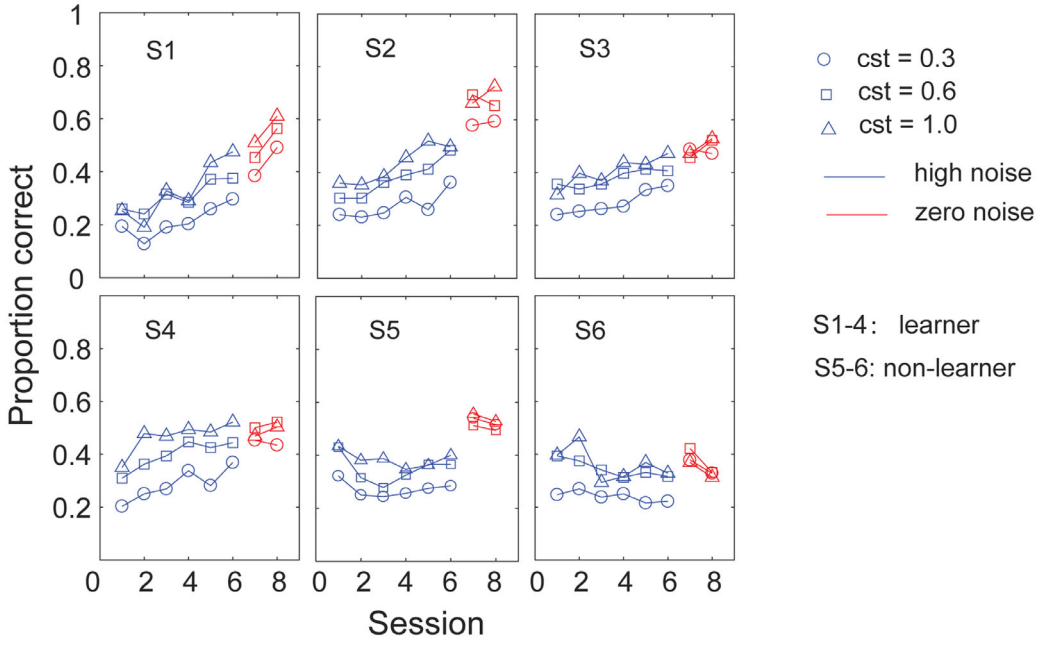

Figure B1. Individual observer data in Experiment 1 (proportion correct version). Observers S1–S4 showed improvement, while observers S5–S6 did not.

blocks within session as the random variable, which is conservative; see Experiment 2 below.) In short, four observers (S1–S4) showed robust learning while the other two (S5–S6) got worse with practice, and these statistical conclusions were consistent with the analysis of the confusion matrix data (see Appendix C). Note that the non-learners here might learn in other or related paradigms.

For each observer, we also fit Weibull functions to the contrast psychometric functions (with three contrasts) for each session to estimate the contrast threshold learning curves (see Experiment 1, main text, for details) ($r^2 = 0.8275 \sim 0.9473$). The resulting learning curves for individual observers are shown in Figure B2 with the best-fitting exponential learning function (S1–S4): $c_\tau(t) = \lambda' e^{-\beta t} + \alpha'$, where $c_\tau(t)$ is the threshold at time (session) t, $\lambda' + \alpha'$ is the initial threshold, α' is the asymptotic threshold after learning, and β is the learning rate. For non-learners (S5 and S6), the system of exponentials defaulted to the mean. The best-fit parameters are listed in Table B1 (top), along with r^2s. We chose the exponential form as it best characterized individual observers in previous investigations (Zhang et al., 2019a; Zhang et al., 2019b; Heathcote et al., 2000; Dosher & Lu, 2007). (Power function fits led to equivalent conclusions; statistically discriminating the exponential and power forms would require more extended training data.)

Changes in the confusion matrices (see Appendix C) provided information unavailable in the mere failure to improve proportion correct: the non-learners shifted tactics towards stereotyped responses clustered around a few responses for low and others for high spatial frequency stimuli. Confusion matrices made it possible to observe non-optimal changes in response patterns even as accuracy of performance was unchanged or even worsening somewhat with additional practice, see Appendix C.

Experiment 2

There was also strong overall evidence in Experiment 2 of learning in spatial frequency identification, as seen in the directly measured contrast threshold learning curves (see Figure B3, main curve. The inset is the proportion correct across sessions for each observer). Some observers (S1–S3) showed clear learning (decrease of thresholds), one (S4) showed learning albeit near ceiling thresholds (together with increases in weighted-κ despite the adaptive staircase efforts to hold accuracy, see the inset of subfigure), whereas others (S5–S7) showed little systematic learning. To assess learning for each observer individually, an analysis of variance was performed on contrast thresholds with blocks within session as the random factor. This may be somewhat conservative, as any systematic effects of block within session appear in the error term. The main effect of training session was significant in three *learners* (S1: $F(5, 15) = 15.11$, $p < 0.0001$; S2: $F(5, 15) = 10.10$, $p \approx 0.002$; S3: $F(5, 15) = 4.56$, p $\approx$ 0. 0.01, respectively); marginally significant in the near-ceiling learner (S4: $F(5, 15) = 2.80$, $p \approx 0.055$); and not significant in the other three (S5: $F(5, 15) = 1.11$ $p \approx 0.396$; S6: $F(1, 15) = 0.73$, $p \approx 0.6137$; S7: $F(1, 15) = 2.21$, $p \approx 0.108$; respectively). Substituting blocks

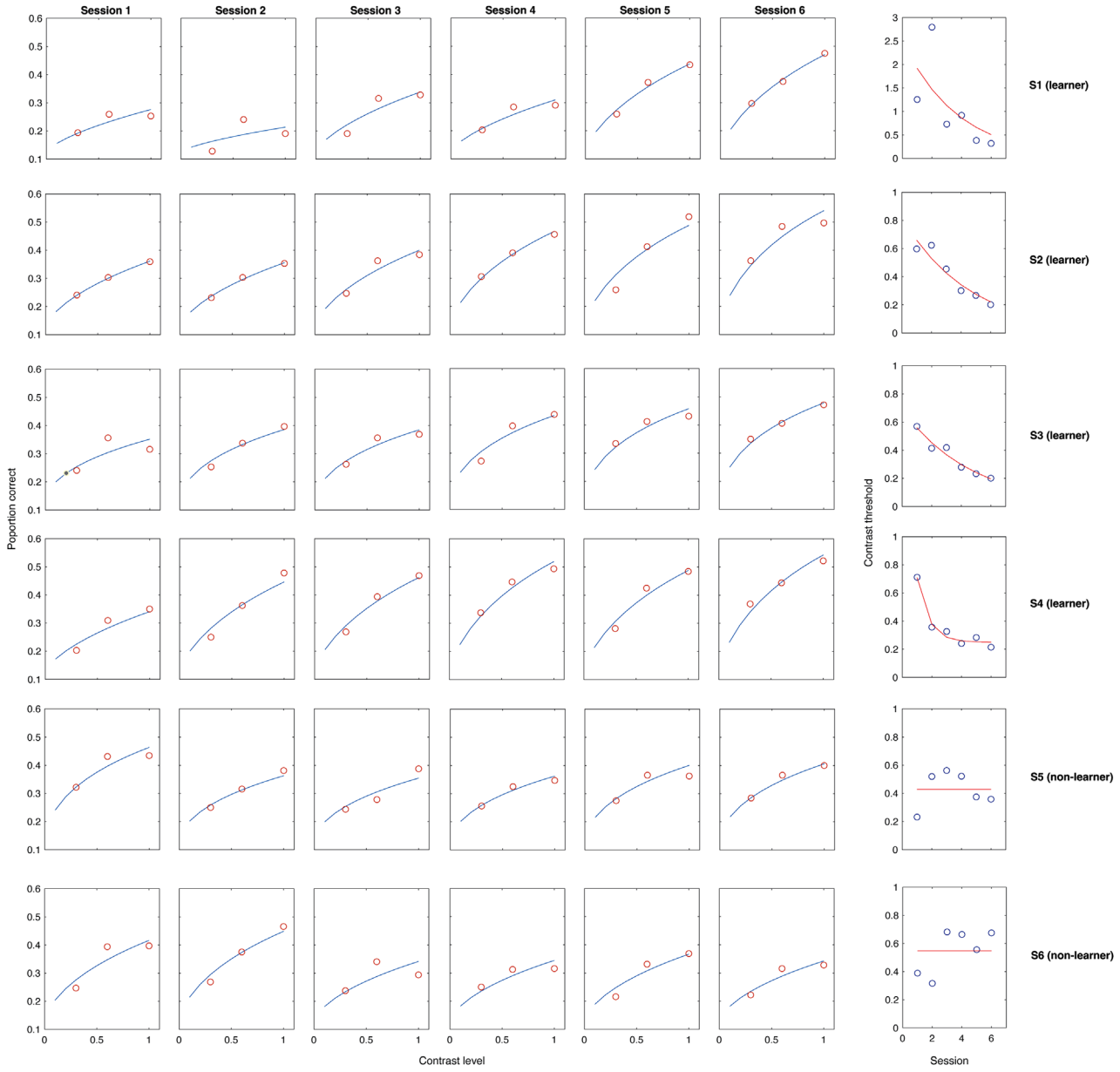

Figure B2. Weibull-fit thresholds from each observer in Experiment 1, and the exponential fit to the threshold learning curves. Thresholds from observers S1–S4 decreased, while thresholds for observers S5–S6 did not.

for observers in the $p_{BIC}(D|H_1)$ computation led to values consistent with these conclusions: S1: $p > 0.999$; S2: $p > 0.999$; S3: $p > 0.999$; S4: $p > 0.992$; S5: $p = 0.071$; S6: $p = 0.001$; S7: $p > 0.934$ (for deterioration rather than improvement in performance for S7). The smooth curves in Figure B3 are the best-fit exponential learning functions to the five sessions of learning in external noise, with parameters and r^2 listed in Table B1 (bottom). (Again, power functions provided a similar analysis, with a reduced form with $\alpha = 0$.) Note that the proportion of four of seven learners (and ceiling learners) was similar to, if slightly lower than, the two-thirds proportion of learners observed in Experiment 1 and the related experiments in our laboratory.

Appendix C: Confusion matrices, weighted-κ, and information transmitted (I_t)

Experiment 1

Confusion matrices reveal the similarity structure limiting absolute identification. As described in the main text, training improved the average confusion matrices as measured by weighted-κ scores. Analyses of variance were performed on the weighted-κ scores for each individual observers (block within session as the random factor). In all observers, the performance

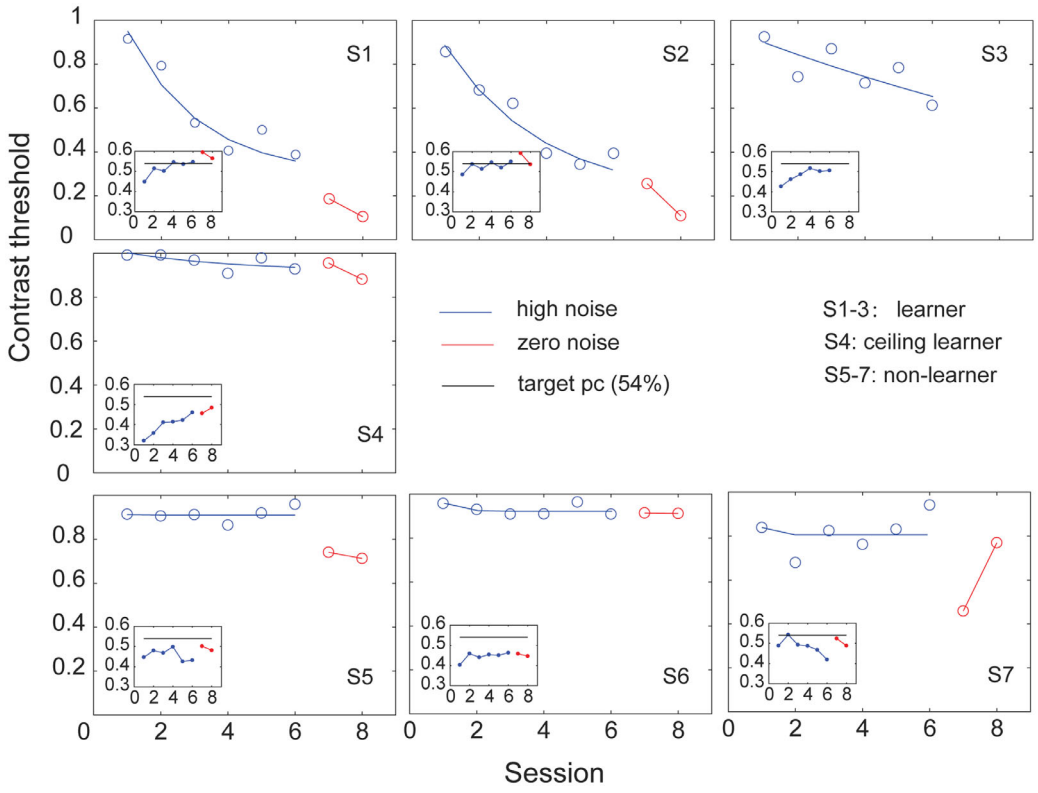

Figure B3. Thresholds and proportion correct (inset) for each observer in experiment 2. Learners, (S1–S3) showed learning in the decrease of thresholds (fit with exponential curves), and accuracy performance that approached the target 54% correct. The ceiling learner (S4) continued over sessions to perform near the threshold ceiling (1.0 contrast), but their performance accuracy improved. The non-learners (S5–S7) also showed closer to ceiling thresholds with little improvement in performance accuracy. Note that while falling short of the target accuracy of the adaptive staircase, the performance nonetheless preserves significant information about the stimulus (see Figure C1 below). See the text for explanations.

Subject	λ'	β	α'	r^2
Experiment 1				
S1 (learner)	2.5069	0.2668	0	0.4096
S2 (learner)	0.8203	0.2185	0	0.9019
S3 (learner)	0.6903	0.2100	0	0.9517
S4 (learner)	1.6585	1.2839	0.2497	0.9715
S5 (non-learner)*	0.4282	–	–	0
S6 (non-learner)*	0.5470	–	–	0
Experiment 2				
S1 (learner)	0.9609	0.4648	0.3347	0.9153
S2 (learner)	0.8979	0.3412	0.2287	0.9336
S3 (learner)	0.9454	0.0582	0	0.5720
S4 (ceiling learner)	0.1049	0.3086	0.9204	0.4283
S5 (non-learner)*	0.9127	–	–	0
S6 (non-learner)	0.9302	–	–	0
S7 (non-learner)*	0.8130	–	–	0

Table B1. Exponential fits to the thresholds of individual subject. (Top: Experiment 1; Bottom: Experiment 2).
*For non-learners, the thresholds did not improve or even deteriorated. The exponential fit was essentially a flat line, which we show as a submodel consisting of the average threshold, listed under λ. Note that the non-learners in these experiments might learn successfully in other paradigms or tasks.

is better for higher contrast levels, but over sessions the weighted-κ scores only improved for the four learners, while showing deterioration for the two non-learners. The statistics are as follows: perceptual learning improved weighted-κ for four observers (S1: $F(5, 15) = 7.727$, $p < 0.0009$ for session, $F(2, 6) = 15.651$, $p = 0.0042$ for contrast, and $F(10, 30) = 3.260$, $p < 0.006$ for the interaction; S2: $F(5, 15) = 13.731$, $p < 0.0001$ for session, $F(2, 6) = 56.238$, $p < 0.0001$ for contrast; S3: $F(5, 15) = 2.254$, $p < 0.10$ for session, $F(2, 6) = 34.413$, $p = 0.0005$ for contrast; S4: $F(5, 15) = 11.151$, $p < 0.0001$ for session, $F(2, 6) = 55.585$, $p < 0.0001$ for contrast). Weighted-κ deteriorated over sessions for two observers (S5: $F(5, 15) = 4.711$, $p < 0.009$ for session, $F(2, 6) = 91.954$, $p < 0.001$ for contrast; S6: $F(5, 15) = 3.949$, $p < 0.02$ for session and $F(2, 6) = 34.462$, $p < 0.0005$ for contrast). The corresponding $p_{BIC}(D|H_1)$ computation were consistent with these results: S1: $p > 0.998$; S2: $p > 0.856$; S3: $p > 0.999$; S4: $p > 0.999$; S5: $p > 0.998$; S6: $p > 0.867$ (for deterioration rather than improvement in performance for S5 and S6). Analyses of information transmitted values (I_t) were similar (because these scores are transformations of proportion correct, on the accurate response diagonal).

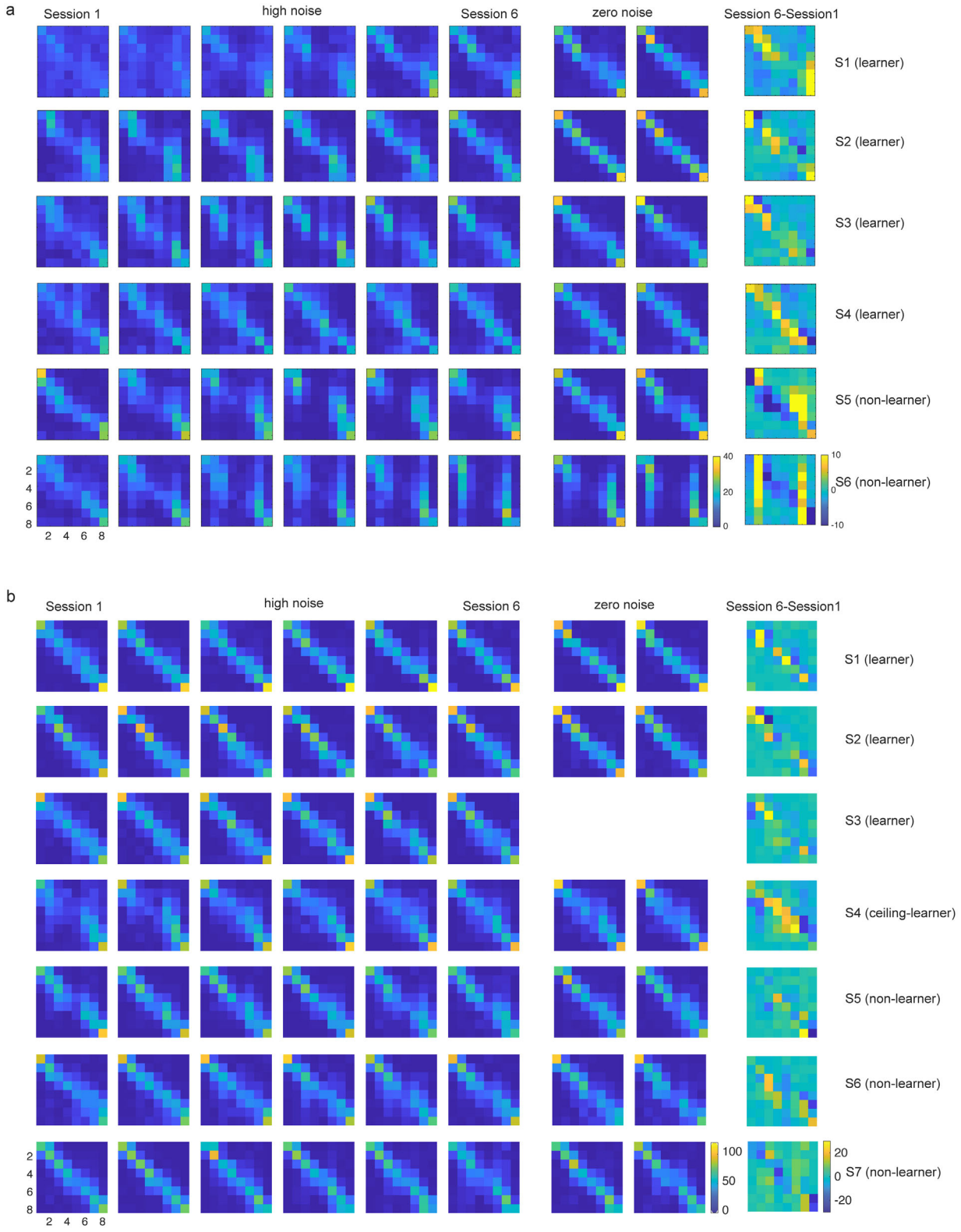

Figure C1. Confusion matrices shown as heatmaps for individual observers. (**a**) In Experiment 1, the confusion matrices showed improvements for learners whereas non-learners developed biases. (**b**) In Experiment 2, consistent with the use of an adaptive staircase to target 54% correct, most confusion matrices remained similar while the ceiling learner showed most improvement.

Examination of the heatmaps for individual observers (Figure C1a) shows that the latter two individuals began with reasonable performance early and for whatever reason developed biases to concentrate responses around one low and one high spatial frequency category label. This is a situation in which the confusion matrices provide information over and above the general accuracy of performance that reveal the underlying behavior.

Finally, the confusion matrices revealed some modest end-anchor effects (bright cells for responses 1 and 8, or the lowest and highest spatial frequency stimuli, respectively) (for other cases in the literature, see e.g. Ward & Lockhead, 1970; Luce et al., 1982). End-anchor effects also emerge from the model, where they emerge naturally from contributions of activation in sensory representations centered beyond the range of the tested stimuli (see model).

Experiment 2

Because, in this experiment, observers were trained at threshold by adaptively changing contrast to achieve the target 54% correct, changes in the confusion matrices were expected to be more subtle. Confusion matrices are shown for individual observers in Figure C1b. These look like the aggregate data for the four learners (S1–S4), but quite different for the non-learners. These latter observers (S5–S7) initially performed the task approximately as well as other observers, and simply did not learn; visual examination of the non-learner heatmaps showed no obvious increase in biases. Changes in weighted-κ were subtle in this experiment due to the adaptive staircase controlling accuracy. As cited in the main text, in the average data weighted-κ to increased from 0.396 to 0.453 over sessions. For individual observers: S1: 0.424 to 0.526; S2: 0.460 to 0.528; S3: 0.393 to 0.483; S4: 0.259 to 0.426; S5: 0.415 to 0.399; S6: 0.361 to 0.434; S7: 0.463 to 0.374. Analyses of information transmitted values (I_t) were similar. The learners (S1–S4) showed the subtle improvements of the aggregate data, as expected. Whether performance in non-learners (S5–S7) reflected a tactical decision to drift or inattention in the more difficult task or inefficient learning at the performance level enforced by the adaptive procedure (54%, or $d' = 1.47$) is not clear.