The Dynamics of Perceptual Learning: An Incremental Reweighting Model

Alexander A. Petrov and Barbara Anne Dosher
University of California, Irvine

Zhong-Lin Lu
University of Southern California

The mechanisms of perceptual learning are analyzed theoretically, probed in an orientation-discrimination experiment involving a novel nonstationary context manipulation, and instantiated in a detailed computational model. Two hypotheses are examined: modification of early cortical representations versus task-specific selective reweighting. Representation modification seems neither functionally necessary nor implied by the available psychophysical and physiological evidence. Computer simulations and mathematical analyses demonstrate the functional and empirical adequacy of selective reweighting as a perceptual learning mechanism. The stimulus images are processed by standard orientation- and frequency-tuned representational units, divisively normalized. Learning occurs only in the "read-out" connections to a decision unit; the stimulus representations never change. An incremental Hebbian rule tracks the task-dependent predictive value of each unit, thereby improving the signal-to-noise ratio of their weighted combination. Each abrupt change in the environmental statistics induces a switch cost in the learning curves as the system temporarily works with suboptimal weights.

Keywords: perceptual learning, cortical plasticity, Hebbian learning, connectionist model, linear classifier

Perceptual learning refers to performance improvements in perceptual tasks as a result of practice or training. Perceptual learning has been of particular interest as it may reflect plasticity at different levels of perceptual analysis—from changes in early sensory representations to higher order changes in the way these representations are used in a task. When the effects of learning are specific to a stimulus characteristic coded early in the visual system, such as retinal location or orientation, an early site of plasticity has generally been inferred (e.g., Karni & Sagi, 1991). In contrast with this position, Dosher and Lu (1998, 1999) proposed that in many cases the behavioral improvement could reflect task-specific learned reweighting of the "read-out" connections from early visual representations, with no changes in the representations themselves. The first view identifies plasticity with changes in the earliest possible areas in the visual cortex; we refer to it as the *representation modification* hypothesis. The alternative *task-specific selective reweighting* hypothesis focuses on the connections from the sensory representations to decision. In this latter view, specificity (lack of positive or negative transfer) can occur if either (a) the sensory representation, (b) the decision structure, or (c) both are distinct between two task or stimulus situations.

Many studies in the prior literature test transfer of perceptual learning between conditions that plausibly involve distinct stimu-

lus representations (e.g., vertical vs. horizontal motion direction). The individual representational units are tuned to specific properties of the stimuli, and the behavioral manifestations of perceptual learning are also stimulus-specific. However, this does not necessarily imply that the behavioral improvement stems from representational improvement (Dosher & Lu, 1998; Mollon & Danilova, 1996). It is entirely possible, indeed probable, that the system learns which features of the redundant, multifaceted early representations are most diagnostic for the task at hand and strengthens the read-out connections from the units encoding these features. The response accuracy improves as the system gradually assigns higher weights to the relevant features and lower weights to the irrelevant ones. As each connection is anchored in a stimulus-specific unit, the effects of the selective reweighting inherit the same specificity.

In this article, we propose a new task analysis that allows us to systematize and understand the prior literature on perceptual learning and transfer. This analytical framework identified critical tests of the representation modification and selective reweighting hypotheses. We report an experiment in which the target stimuli—and, hence, the signal representation—are identical, while a background noise *context* differs between two task environments. The observers perform the same orientation-discrimination task throughout 8 days of training. The specificity manipulation is implemented in the noisy contexts mixed with the task-relevant targets—two such contexts alternate in an *A-B-A-B* design. We observe significant switch costs (interference) after each change in context. Moreover, the magnitude of the cost remains undiminished for as long as five switches and 9,600 trials, indicating that a single set of connections is involved in both contexts. The recurring cost pattern mirrors the nonstationary structure of the task environment and strongly suggests that perceptual learning is associative and statistically driven.

The existence of (negative) transfer between task environments with identical target stimuli might seem to favor a form of per-

ceptual learning that is based on representation modification. Instead, we show that an incremental Hebbian reweighting network model can account quantitatively for the complex pattern of learning and switch costs in our nonstationary training protocol. The model develops the multichannel perceptual template model proposed by Dosher and Lu (1998, 1999) and extends it with a fully functional representation subsystem and a biologically plausible learning rule. The model takes grayscale images as inputs and improves its accuracy on a trial-by-trial basis. The model's performance is directly comparable to the human data, which is rare in the perceptual learning literature. Computer simulations demonstrate the model's ability to account quantitatively for the learning dynamics across three target contrast levels, the context-dependent switch costs, and several other patterns in a complex data set. Mathematical analyses show how the switch costs predicted by the model are direct consequences of the different predictive value of certain stimulus features in the two contexts. Statistical analysis of the stimulus space shows that this perceptual learning task approximates a linear boundary problem with different optimal boundaries in different environments. The model generates new testable predictions, and the task analysis suggests a new classification of perceptual tasks.

Together, the empirical results and the model provide an existence proof that the task-specific reweighting hypothesis is sufficient to account naturally and quantitatively for a challenging set of perceptual learning phenomena.

Representation Modification and Task-Specific Reweighting

Task Analysis

Even the simplest task requires a system with at least three components: a stimulus representation area, a decision unit, and a connection pathway between them. Perceptual learning can be accomplished through representation modification, selective reweighting, or both. The observed specificity to basic stimulus features such as retinal location or orientation requires only that one end of each modifiable connection is attached to a unit characterized by that specificity. Both representation modification and task-specific reweighting are consistent with that requirement. In many cases, the two hypotheses make similar predictions for the pattern of performance in a *transfer* task following a *training* task. In other cases, the two hypotheses make distinct predictions, depending on the nature of the task comparisons. Here, we provide a new task-analysis framework that allows us to classify and interpret the prior literature and identify what critical tests can discriminate between perceptual learning models.

Four key stimulus and task situations probing perceptual learning are illustrated schematically in Figure 1. In the first situation (see Figure 1A), the training and transfer environments activate distinct sensory units and distinct weighted connections from those units to the decision units. In this situation, training in one environment should transfer little to the other environment regardless of whether perceptual learning occurs through representation modification or through reweighting. If the two environments alternate multiple times, the performance in each of them should improve independently from any intervening training in the other. Both hypotheses predict independent learning specific to each task.

In the second situation (see Figure 1B), the task-dependent decision structure is shared, but the stimulus representations still
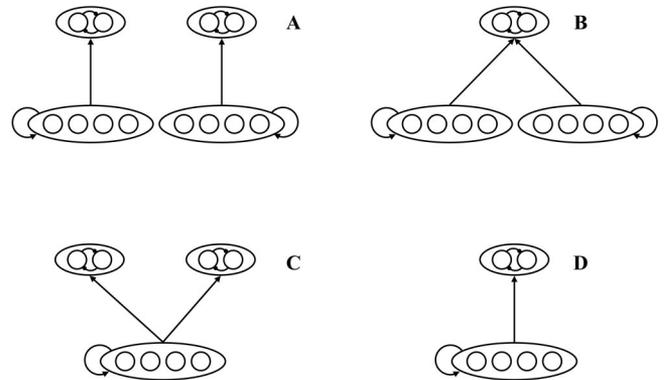

Figure 1. When an observer trains a perceptual task and is subsequently tested with a second task and/or stimuli, different patterns of overlap can occur between the neural structures engaged during training and test. The two tasks can depend on same (B, D) or different (A, C) response-selection areas, and the two stimulus environments can activate overlapping (C, D) or nonoverlapping (A, B) neuronal populations in the sensory areas. The implications for perceptual learning are discussed in the text.

do not overlap during training and testing. For instance, this configuration is likely when the same stimuli are presented first in one visual hemifield and then in the other. Again, both hypotheses predict independent learning of the two tasks. Outcomes other than substantial specificity or independence of learning in Situations 1A or 1B would be surprising, and independence of learning cannot distinguish representation enhancement and reweighting.

In the third situation (see Figure 1C), the sensory representation is shared, but the weighted connections from the (shared) sensory units to the decision units are separate for the two tasks. The two hypotheses make different predictions in this case. If perceptual learning involves representation modification, training on the first task must impact performance on the second. If learning instead involves reweighting of the read-out connections, performance in the two tasks may still be independent from initial training to transfer or across multiple alternations of training.

In the final situation (see Figure 1D), both the stimulus representation and the task-dependent decision structure are shared between the two environments. One way this can be arranged is by keeping the target stimuli and the task constant and manipulating only the task-irrelevant background context. With such significant overlap in the neuronal substrate, neither hypothesis predicts complete specificity, but the situation is now constrained enough to make finer-grain predictions possible. In the unitary system in Figure 1D, the detailed dynamics of learning in nonstationary environments becomes very informative. Tracking the learning curves across multiple switches (rather than a single posttest) can reveal the underlying plasticity. The interaction between the two environments could take the form either of enhanced cross-transfer or, more likely, switch costs.

Our approach in this article is to develop a specific quantitative model based purely on reweighting and to test it on a rich, dynamic, and highly constrained data set that could plausibly reflect representational change. The overall goal is to evaluate the empirical adequacy and plausibility of the task-specific reweighting hypothesis. First, however, we evaluate each hypothesized form of learning in light of the existing behavioral and physiological evidence.

Psychophysical Evidence: Specificity

The behavioral specificity of perceptual learning (e.g., Ahissar & Hochstein, 1993, 1996; Ball & Sekuler, 1987; Crist, Kapadia, Westheimer, & Gilbert, 1997; Fahle, 1997; Fiorentini & Berardi, 1980; Shiu & Pashler, 1992) has served as one basis for claims of plasticity in primary visual cortex, where plasticity has often, by hypothesis, been assigned to the earliest level exhibiting a cellular basis consistent with the specificity (e.g., Karni, 1996). Specificity for the eye of training, in particular, has been interpreted as evidence of plasticity within the primary visual cortex (V1), the only cortical site where the input from the two eyes is still segregated (Karni & Sagi, 1991). Specificity of learning has also been observed for visual field location, orientation, spatial frequency, size, and motion direction (see Ahissar & Hochstein, 1998; Fahle & Poggio, 2002; Gilbert, Sigman, & Crist, 2001, for reviews). Not all forms of specificity hold for all paradigms, and there is always at least partial transfer of learning, especially interocular transfer (Ahissar & Hochstein, 1996; Lu, Chu, Lee, & Dosher, 2005; Schoups & Orban, 1996). Still, the general phenomenon is well established.

The question is which task situation (see Figure 1) provides the best description of the tasks in the literature, and what can be concluded about the locus of perceptual learning in each particular study.

Most behavioral experiments use stimuli and/or tasks that are sufficiently different to induce unambiguous specificity of the learning effects during training and in a subsequent transfer test. Thus their underlying structure can be reasonably approximated by either Figure 1A or 1B (or some intermediate version with partial overlap between the stimulus representations). Although these experiments have revealed many interesting and important perceptual learning phenomena, their designs are not constraining enough to discriminate the possible learning mechanisms. The observed specificity is equally consistent with representation modification, task-specific reweighting, or some combination of the two.

For example, in one prominent paradigm, improvements in perceptual judgments with stimuli of a particular orientation tend to transfer very little when the trained observers are asked to perform the same task at the orthogonal orientation (Ahissar & Hochstein, 1996, 1997; Crist et al., 1997; Fahle & Edelman, 1993; Fiorentini & Berardi, 1980; Karni & Sagi, 1991; Poggio, Fahle, & Edelman, 1992; Schoups, Vogels, & Orban, 1995; Vogels & Orban, 1985). By the standard inference (the minimal level hypothesis; Karni, 1996), specificity to orientation and/or retinal location is attributed to cortical areas with orientation selectivity, retinotopic organization, and relatively small receptive fields: V1, V2, or possibly V4. However, orthogonally oriented stimuli activate largely nonoverlapping neuronal populations in these cortical areas. Thus, even though the same brain area is involved in both cases, the representations are effectively disjoint as depicted in Situations A and B in Figure 1. These are precisely the situations where both hypotheses predict that learning one orientation should transfer relatively little to the other. *Stimulus specificity* in this and a number of other classic paradigms is not a diagnostic observation with respect to the distinction between representation change (early locus) and reweighting.

Another prominent set of studies uses identical stimuli but different tasks and argues for *task specificity* of perceptual learning. In Ahissar and Hochstein (1993) for instance, the stimuli were rectangular arrays of short oriented line elements with a single element of a different orientation; the two tasks were (a) to locate the odd element or (b) to classify the long dimension of the rectangular array. Performance on both tasks improved with practice but there was little or no transfer between them, even when the two attributes were perfectly correlated during training (Ahissar & Hochstein, 1993). Although the stimuli are identical for the two tasks, the discriminative stimulus information is not and the very different task demands make it likely that the two tasks use independent sets of connections from a shared representation (see Figure 1C) or that the (functional) representations are distinct (see Figure 1A). Assuming the stimulus representations do overlap, these results clearly suggest a reweighting learning mechanism.

Several other empirical examples are similarly compatible with the reweighting hypothesis (Dosher & Lu, 1998, 1999, 2005). Fahle (1997) tested curvature, orientation, and vernier discrimination. As all three tasks are thought to rely on related orientation coding mechanisms (Wilson, 1986), this study plausibly did test common representations and found no transfer of learning. In another strong data set, orientation discrimination was nearly independent (< 10% reduction) of dramatic improvements (65%) in vernier training (Crist et al., 1997). These cases in the literature are also, presumably, examples of common representations but independent sets of read-out connections (see Figure 1C).

In summary, the psychophysical evidence advanced in support of the representation modification hypothesis seems equally consistent with the selective reweighting hypothesis. In many cases, the training and transfer tests do not share stimulus representations and hence are nondiagnostic examples of specificity (see Figures 1A–1B). In other cases, the same or similar inputs are likely to be involved but the existence of considerable task specificity suggests that nonoverlapping sets of connections may be involved, perhaps reflecting significantly different tasks (see Figure 1C). The results in these putative cases of 1C are apparently consistent with reweighting. In this article, we designed a potentially diagnostic example in which two task variants depend upon the same representations and the same task-dependent decision system (see Figure 1D). This is a case in which task interaction is expected and representation modification is a possible mechanism to account for perceptual learning.

Physiological Evidence: Sensory Representations

Plasticity of early sensory representations has been documented in a range of invasive neuroscience protocols (see Buonomano & Merzenich, 1998; Das, 1997; Gilbert et al., 2001, for reviews). For example, following retinal lesions, the receptive fields of deafferented neurons expand to include surrounding regions of the visual field (Chino, Kaas, Smith, Langston, & Cheng, 1992; Darian-Smith & Gilbert, 1994; Eysel, 2002; Gilbert & Wiesel, 1992; Kaas et al., 1990; Schmid, Rosa, Calford, & Ambler, 1996). These results establish the potential for continued plasticity and recruitment in the adult visual cortex.

Do similar modifications occur under noninvasive training conditions in vision? Although evidence for redistribution of cortical territory has been reported in other sensory modalities, notably somatosensory (Jenkins, Merzenich, Ochs, Allard, & Guic-Robles, 1990; Jones, 2000; Recanzone, Merzenich, Jenkins, Grajski, & Dinse, 1992) and auditory (e.g., Recanzone, Schreiner, & Merzenich, 1993; Weinberger, Javid, & Lepan, 1993), there is little

compelling evidence for long-term cortical reorganization of the earliest visual representations (V1, V2) induced solely by behavioral training in adult animals.

Three recent studies reported single-cell recordings in early visual areas following extensive practice in adult monkeys (Crist, Li, & Gilbert, 2001; Ghose, Yang, & Maunsell, 2002; Schoups, Vogels, Qian, & Orban, 2001). All three experiments revealed the same overall pattern: Behavioral measures showed marked improvements in performance, and yet, basic receptive field properties, such as location, size, and orientation selectivity, were largely indistinguishable between the trained and untrained regions in both V1 and V2. Ghose et al. (2002) performed correlation analyses on eight independent receptive field parameters in four neuronal populations with respect to distance from the training stimuli and preferred orientation. Only one of the 64 analyses was significant, and it involved an untrained population. Schoups et al. (2001) also found that the peaks of the tuning curves were evenly distributed over all orientations both in trained and untrained neurons, although the slopes of the tuning curves were increased for trained neurons with preferred orientations about 20° away from the trained orientation. However, these changes were modest and not sufficient to fully explain the behavioral improvement from training. These results were not replicated by Ghose et al. (2002), who found no significant changes in the slope of the tuning curve. The average peak slopes (and the tuning bandwidths) did not vary between trained and untrained orientations in their sample. If anything, training appears to be associated with a slight decrease in the number of neurons whose preferred orientation is near to the trained orientation in these regions of visual cortex (Ghose et al., 2002, Figure 4). Crist et al. (2001) constructed topographic maps of the centers of single-unit receptive fields in two trained monkeys. The magnification factors across both trained and untrained hemispheres were almost identical to one another and to measurements in untrained monkeys. In contrast, Yang and Maunsell (2004) reported that massive practice in match-to-sample orientation discrimination did produce some modest sharpening of the tuning curves in intermediate visual cortex (V4) for selected cells responsive to the trained orientation. The changes were not, however, sufficient to account for the dramatic behavioral improvement without additional assumptions of selective read-out consistent with the reweighting hypothesis.

In all these physiological recordings from primary visual cortex, learning does not appear to be driven by neuronal recruitment, increased cortical territory, or other major restructuring of the early representations. So far, the evidence for representation modification consists of subtle changes in the tuning properties of selected neurons in V4 and conflicting reports of changes in V1 and V2. All these changes can be (and perhaps even must be) coupled with selective reweighting of the read-out connections to account for the behavioral improvement. Of course, it is also possible that the effects of training cannot be traced in the static properties of the representations but occur only when the monkey actually carries out a specific task (Li, Piëch, & Gilbert, 2004). This latter proposal of "multiplex" representations may be isomorphic with task-specific reweighting having independent weight structures for different tasks and is further considered in the General Discussion.

Finally, there is some evidence from two neuroimaging reports of practice-induced changes in early retinotopic cortex (Schiltz et al., 1999; Schwartz, Maquet, & Frith, 2002). This evidence, however, is affected by several challenges to interpretation. Neuroim-

aging studies of perceptual learning are methodologically complicated by the inherent confound between practice and overall performance level. As a compromise, the stimulus intensity is chosen to yield performance levels either close to chance (53%–58% correct; Schwartz et al., 2002) or close to ceiling (94%–98% correct; Schiltz et al., 1999). Schwartz et al. (2002) measured the blood oxygenation level-dependent (BOLD) response in functional MRI after a single session of monocular training in visual texture discrimination. They found an increased response in early retinotopic cortex (right calcarine sulcus) to stimuli presented to the trained eye relative to stimuli presented to the untrained eye. Schiltz et al. (1999) measured the regional cerebral blood flow by means of positron emission tomography (PET) before and after 10 sessions of training in orientation discrimination. There was an orientation-specific decrease in the blood flow to striate and extrastriate cortices in the posttraining scan relative to the pretraining scan. There was also a nonspecific decrease in cerebellar blood flow (Schiltz et al., 1999; see also Vaina, Belliveau, des Roziers, & Zeffiro, 1998). More research is needed to clarify these contradictory findings—an activation increase in functional MRI and decrease in PET.

Selective Reweighting

We conclude from this analysis of the behavioral and physiological literature that the selective reweighting hypothesis provides an explanation of perceptual learning that is both intuitively plausible and consistent with the available evidence. Even if representation modification also plays a significant role, it seems inevitable that reweighting will still be required to fully account for behavior. Moreover, it is quite possible for selective reweighting to operate on fixed representations, but it is hard to imagine representation modification without reweighting. Task-dependent selective reweighting thus emerges as a fundamental perceptual learning mechanism that is well worth studying in detail.

The selective reweighting explanation for perceptual learning was initially proposed by Dosher and Lu (1998, 1999). They suggested that perceptual learning could be accomplished by selective reweighting within a multichannel observer model. The channels in the generic multichannel model have templates tuned to different properties of the input image and compute a noisy coarse-coded representation of the stimulus. A weighted average of the noisy channel outputs is then submitted to a decision process within a signal detection framework (e.g., Macmillan & Creelman, 1991).

Dosher and Lu's (1998, 1999) proposal was based on an analysis of performance improvements within the context of a (signal detection) model of the observer (Lu & Dosher, 1999). The observer analysis estimates different internal noise and inefficiencies—and how they change with perceptual learning—on the basis of observer performance in titrated external noise tests. Characterizing observer limitations in equivalent noise terms has provided important insights into fundamental properties of the perceptual system (e.g., Burgess, Wagner, Jennings, & Barlow, 1981; Lu & Dosher, 1999). One important property of the perceptual system relevant for the present analysis is system nonlinearity and gain control. Performance must be measured for at least three criterion performance levels to specify system nonlinearity (Dosher & Lu, 1999; Lu & Dosher, 1999) in an observer analysis. Performance in our experiment is measured at three levels of physical contrast of

the stimulus. Experimental grounding in this physical manipulation is one important aspect of model testing and analysis.

From a computational point of view, perceptual learning is a statistically driven process of extracting regularities from the stream of experience and capitalizing on them to improve performance. The statistics of the task environment play a critical role in both enabling and shaping this process. Perceptual tasks often require a large sample of experiences, accrued over a large number of trials, to separate the reliable regularities from the task-irrelevant variations. The extended temporal dynamics provides valuable information about the plasticity mechanisms. *Nonstationary environments* whose statistical properties change with time can be particularly informative because they act as "moving targets" and are especially challenging for most learning systems. Indeed, the present experiment uses repeated alternation between two test contexts to reveal the properties of perceptual learning. The longer alternations can be especially important in distinguishing between independent learning in two task environments and constant readjustment of a single learning process as the statistics of the test environment change (Petrov & Anderson, 2005).

In the present article, we report an experiment in which the background noise always has the same contrast but its spectral properties alternate in a nonstationary sequence, thereby highlighting the dynamic aspects of perceptual learning. Three values of contrast are used to allow estimation and testing of system nonlinearity and to ground the performance in a physical stimulus manipulation. Extended task alteration is used to test the response to a nonstationary task environment but also to evaluate the evidence for independent learning versus ongoing adjustments and switch costs.

We also fully implement and test the schematic multichannel reweighting model outlined by Dosher and Lu (1998) and extend it with a biologically plausible, incremental reweighting mechanism to account for the detailed temporal dynamics of learning and specificity. The multichannel representation is implemented by orientation- and frequency-selective filters that process the grayscale stimulus image and are subject to nonlinear normalization and internal noise. An incremental Hebbian learning rule updates the weights of the connections to a decision unit. The resulting fully functional model provides an elaborate and explicit test of the reweighting hypothesis on a complex data set in a nonstationary training environment.

Experiment

The present experiment is explicitly designed to ensure maximally overlapping representations as depicted in Figures 1D or 1C, an extended nonstationary stimulus presentation protocol, high levels of external noise, and multiple target contrast levels. According to the preceding theoretical analysis, this design is well suited to evaluate representation modification and selective reweighting as possible learning mechanisms. It also provides multilayered quantitative data for rigorous model testing.

The observer's task and the target images are fixed throughout 8 days of training—orientation discrimination of *Gabor patches* (windowed sinusoidal gratings) either 10° or −10° from vertical. The nonstationary manipulation involves the context of distracting external noise that surrounds the target (see Figure 2). The noise is filtered to form textures with predominantly left orientation in

contextL and predominantly right orientation in contextR. An L-R-L-R schedule alternates the two contexts in 2-day intervals.

Because the context varies in the same dimension as the target Gabors in this case, the composite stimuli can be either congruent or incongruent. A congruent stimulus is one in which the Gabor and background orientations have the same sign; in an incongruent stimulus, they have opposite signs (see Figure 2). To constrain estimates of the nonlinearity in the perceptual system (Lu & Dosher, 1999), the contrast of the Gabor target is varied to create three difficulty levels. This contrast manipulation provides additional strong constraints on the model because it must account for three parallel learning curves with a single set of parameters based solely on the physical properties of the stimuli themselves.

Method

Stimuli. Each stimulus consists of a Gabor patch $G(x, y)$ embedded in a larger field of filtered visual noise $N(x, y;$ Equations 1 and 2). The observers are instructed to ignore the background and indicate the orientation θ of the Gabor target, which can be either $-10°$ or $10°$ from vertical, as illustrated in Figure 2. Each noise field is filtered so that its spectral power is concentrated in a relatively narrow cone of orientations around a *predominant (midline) orientation* $\phi = -15°$ in contextL and $\phi = +15°$ in contextR.

The luminance $L(x, y)$ of each pixel is an additive mixture of a Gabor term $G(x, y)$ and noise $N(x, y)$, where L_0 is the midgray value of the monitor:

$$L(x, y) = [1 + c_p G(x, y) + c_n N(x, y)]L_0 \qquad (1)$$

$$G(x, y) = e^{-(x^2+y^2)/2\sigma^2} \sin[2\pi f(x \cos\theta + y \sin\theta)]. \qquad (2)$$

The peak target contrast c_p is set to 0.245, 0.160, or 0.106; the peak noise contrast is always the same, $c_n = 0.667$. The sine-phase Gabor patches have spatial frequency $f = 2$ cycles per degree, and the standard deviation of their Gaussian envelope is $\sigma = 0.4°$.

To generate the background $N(x, y)$, an isotropic field of Gaussian noise is filtered in the Fourier domain with the conical filter defined by Equation 3:

$$H_\phi(f_x, f_y) = \left[1 + \frac{(f_x \cos\phi + f_y \sin\phi)^2}{\beta^2(f_x \sin\phi - f_y \cos\phi)^2}\right]^{-1}. \qquad (3)$$

Its cross-section at any spatial frequency is a Butterworth bandpass filter of order 1 (Gonzalez & Woods, 1992) with half-amplitude half-bandwidth $\beta = 0.20 \approx \tan 11.3°$. The filter H_ϕ preserves the spectral power near the peak orientation ϕ and attenuates the orientations progressively away from it (see Figure 3). To generate a stimulus, the algorithm generates a 64×64 sample of iid Gaussian noise and applies a filter tuned for $\phi = -15°$ in contextL and $\phi = 15°$ in contextR. The resulting matrix $N(x, y)$ is then rescaled linearly to zero mean (midgray) and peak contrast $c_n = 0.667$. The standard deviation of the term $c_n N(x, y)$ in Equation 1 varies in the 0.17–0.19 range for most noise patches. Finally, the Gabor $G(x, y)$ is superimposed, and the resulting image is quantized to 256 grayscale levels and clipped within a circular window with radius 32 pixels (≈ 1.44 degrees of visual angle; see Figure 2).

Fourier analysis of the resulting stimuli verifies that they all occupy the same region in the spatial frequency domain and, hence, presumably activate a common neuronal population (cf. Figures 1C–1D). The spectral power within this common region is distributed in different patterns depending on congruence, target contrast, and noise fluctuations. Figure 3 plots representative cross-sections of the power spectra of congruent (left panels) and incongruent stimuli (right panels) in contextL. The spectra in contextR are simply mirror images. The top two panels plot cross-sections at the spatial frequency of the Gabor target (2 cyc/deg). Note the peaks at orientation $\theta = -10°$ for congruent (left) and 10° for incongruent (right)
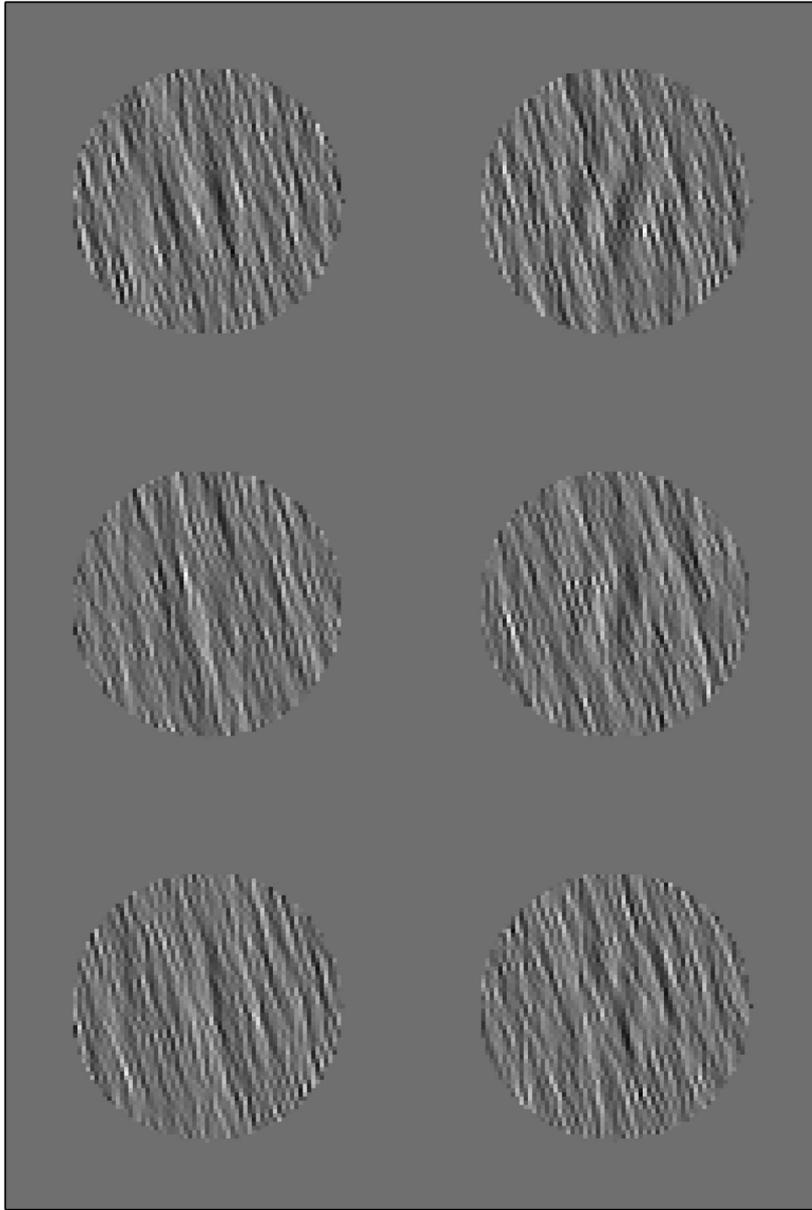
Figure 2. Examples of congruent stimuli (left column, $\theta = -10°$) and incongruent stimuli (right column, $\theta = 10°$). Target contrast decreases from top to bottom. The predominant background orientation is $\phi = -15°$ in all cases.

targets. The impact of the target, however, does not extend to other frequencies as evident from the profiles at 4 cyc/deg (Figure 3, bottom). These profiles are nearly identical, as both are equal to the cross-section of the filter. Note the consistent peak at $\phi = -15°$ and the multiplicative variability profile marked by the dashed lines.

Apparatus. All stimuli were generated in MATLAB (The MathWorks, 1999) in real time and presented with the Psychophysics Toolbox extensions (Brainard, 1997). They were displayed on a NANAO Technology FlexScan 6600 monochrome monitor with P4 phosphor and a refresh rate of 120 frames/sec driven by the internal video card of a Power Macintosh 7300. A special circuit combined two 8-bit output channels of the video card to produce 6,144 distinct gray levels (Pelli & Zhang, 1991). Luminance calibration was performed with psychophysical matching judgments (Lu & Sperling, 1999) and by measurement with a Tektronix Lumacolor J17 photometer. A linear lookup table divided the entire dynamic range of the monitor (from 1 cd/m² to 30 cd/m²) into 256 evenly spaced levels. The background was set at $L_0 = 15$ cd/m². All displays were viewed binocularly with the natural pupil at a viewing distance of approximately 72 cm. The only source of light in the room was the monitor.

Observers. Thirteen paid volunteers participated in the study. All had normal or corrected-to-normal vision.

Design. The learning dynamics was tested in an *A-B-A-B* design. Two groups of observers differed with respect to the context they first trained on: 7 observers began in context*L,* and 6 began in context*R.* The presentation schedule was organized in *blocks* and *epochs.* The context was stationary within each epoch and alternated between them to produce a nonstationary overall environment. There were 8 sessions on separate days, with 4 blocks per day and a total of 32 blocks. The presentation schedule
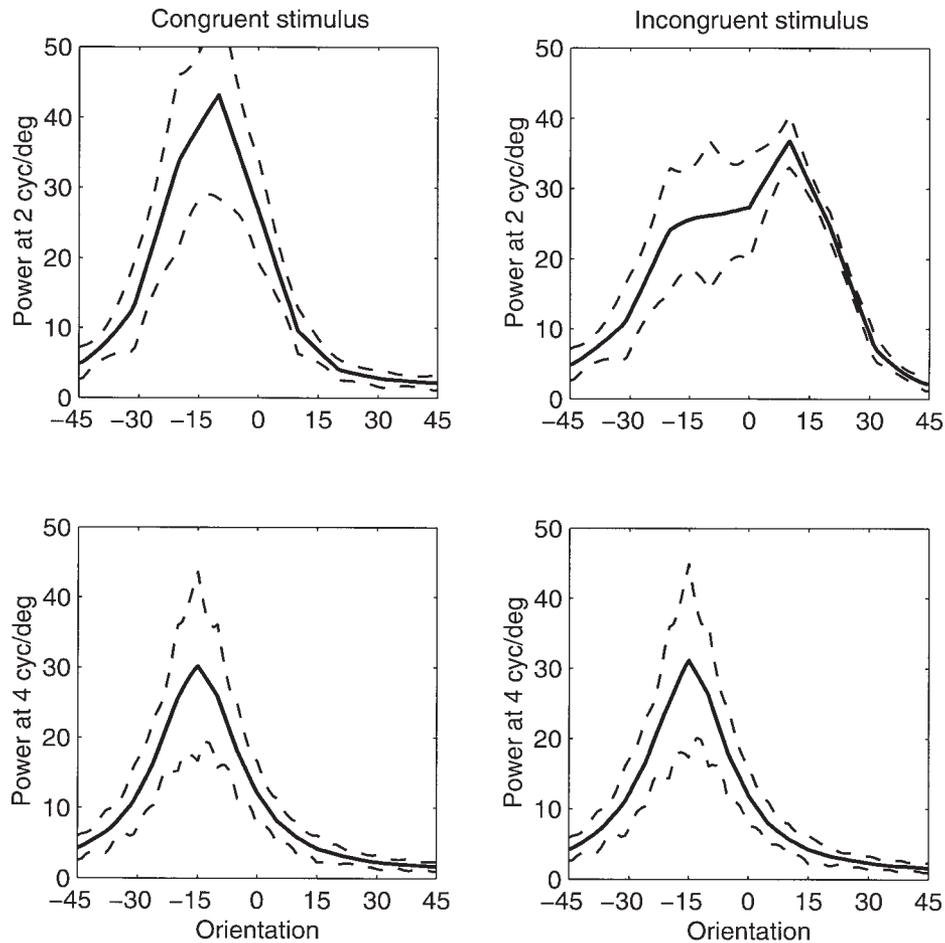
Figure 3. Fourier spectra of congruent (left panels) and incongruent (right panels) stimuli of intermediate target contrast in context*L* (predominantly left orientation). The top row plots cross-sections at the Gabor spatial frequency (2 cyc/deg); the bottom at 4 cyc/deg, which contains only noise. The dashed lines mark ±1 standard deviation from the mean.

was *L-8R-8L-8R-6L-R* for Group 1 and *R-8L-8R-8L-6R-L* for Group 2. All context switches occurred in midsession, avoiding the potential confounds of overnight consolidation or forgetting. Each block consisted of 300 trials in an orthogonal factorial design: 2 Gabor orientations × 3 Gabor contrasts × 2 retinal locations × 25 replications of each stimulus type. The presentation sequence was randomized within each block, and a fresh patch of filtered noise was generated on each trial.

Procedure. Each session began with a few demonstration trials (10 trials on Day 1, 2 on subsequent days) followed by 4 blocks of 300 experimental trials. The observers were instructed to ignore the background and to indicate the orientation of the Gabor target by pressing a key with their left or right hand. Trials with invalid responses were repeated at the end of the block to ensure that there were exactly 1,200 valid observations per session. This occurred on less than 0.01% of the trials. Brief rest periods were allowed between blocks, although many participants chose to skip most of them.

Each trial began with a brief beep and a fixation cross in the middle of the screen. The stimulus appeared for 75 ms at one of two equiprobable locations centered 5° either above or below fixation. Then the screen was cleared, and the observer's response was recorded. Auditory feedback (a "buzz" tone) marked incorrect responses. The next trial began after a 750-ms intertrial interval.

Dependent variables. Probability correct is tabulated separately for congruent and incongruent trials at each contrast level in each block. This

yields 192 data points per participant, each estimated from 50 observations counterbalanced across the two retinal locations. The normal transformation $z = \Phi^{-1}(p)$ linearizes the dependent variable and allows averaging. Extreme frequencies ($p = 50/50$; 29 cases in all) are transformed to $z = 2.33 = \Phi^{-1}(0.99)$. Discriminability measures (d's) are computed by adding the corresponding z values for congruent and incongruent stimuli.

Results and Discussion

The two observer groups, those starting in context*L* and those starting in context*R*, are statistically indistinguishable, $F(1, 11) < 1$, and are combined in a single sample.

Learning dynamics. Figure 4 plots the d' curves at the three difficulty levels, averaged across observers. There is clear evidence of perceptual learning: d' increases substantially over the 8 days (32 blocks) of training. Predictably, the higher the target contrast c_p, the better the discriminability.

The nonstationary context manipulation is implicit in the time variable. The connected lines in Figure 4 belong to the same context; the discontinuities mark context switches. The data suggest transient reductions in performance after each switch. These switch costs indicate that part of the improvement is specific to the
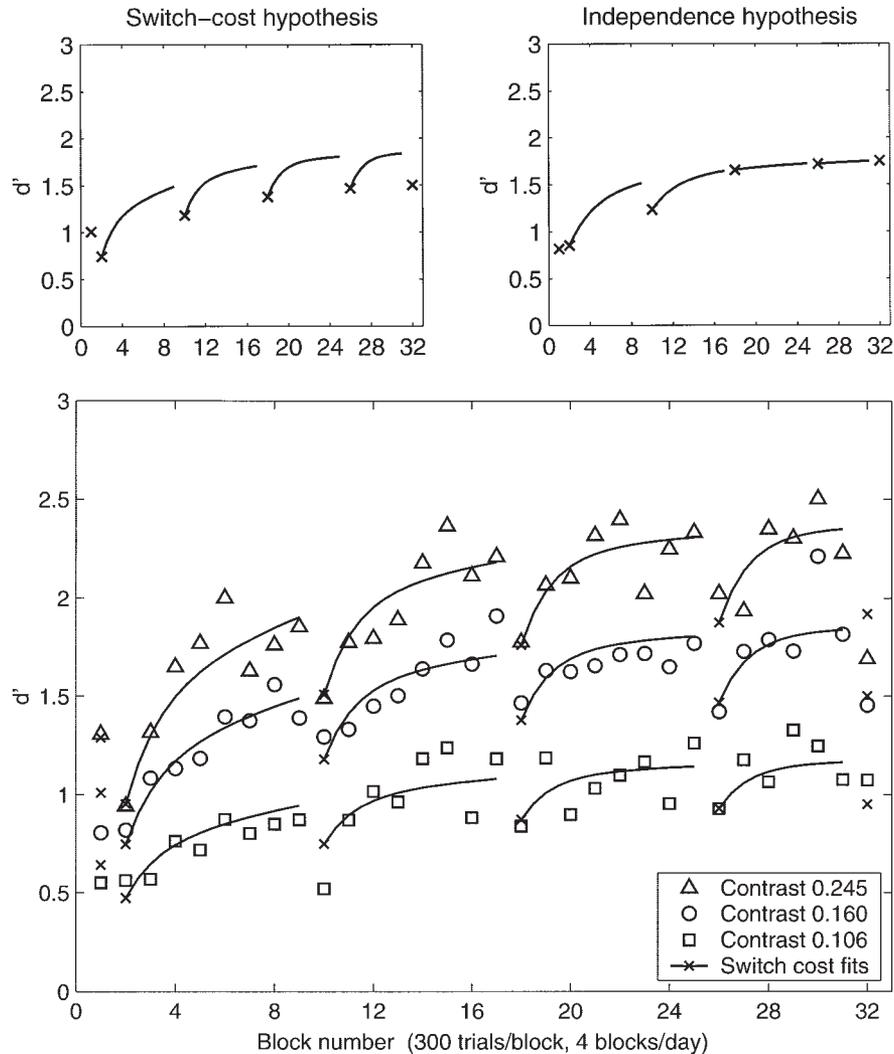
Figure 4. Top panels: Predictions of the two hypotheses discussed in the text (Equation 4). The postswitch blocks are indicated by ×. Bottom panel: d' learning curves for the three target contrast levels (95% confidence interval = ± 0.235). The connected lines belong to the same context; the discontinuities mark context switches. Note the recurring *switch costs*.

background context, replicating the general phenomenon of stimulus specificity of perceptual learning. They are also in accord with a recent study of task-irrelevant learning (Watanabe, Náñez, & Sasaki, 2001).

The *partial context specificity* manifested in the switch costs seems extremely problematic for the representation modification hypothesis. Given the fixed Gabor targets and the extensive representational overlap in both contexts, it is not at all obvious how neuronal recruitment or tuning-curve sharpening would be capable of producing specificity in this instance. The selective reweighting hypothesis, on the other hand, requires switch costs in nonstationary environments, for statistical reasons discussed at length later.

The critical question in the present experiment is whether the costs persist for indefinitely many switches or subside after the first two. If the behavioral improvement stems from selective reweighting of a single set of connections as schematized in Figure

1D, training in one context is expected to interfere with earlier training in the other context. This *switch-cost hypothesis* predicts the seesaw pattern illustrated in the top left panel in Figure 4. On the other hand, if two independent sets of connections are involved (see Figure 1C), the learning curve in each context is expected to increase monotonically regardless of any intervening training. This *independence hypothesis* also predicts a d' decrement after the early switches, in agreement with the default expectation for stimulus-specific perceptual learning. However, it predicts that the d' decrements should diminish rapidly over subsequent switches because the independent learning curves would reach the region of diminishing returns in each context (see Figure 4, top right).

We compare these two hypotheses quantitatively with the aid of the regression model in Equation 4. The learning curve for target contrast i is decomposed into a context-general and a context-specific component relative to the corresponding asymptotic level D_i:

$$d'_i(T, t_s) = D_i(1 - g \, e^{-T/\tau} - s \, e^{-t_s/\tau_s}). \qquad (4)$$

The functional form of the regression equation is the same for both hypotheses; they differ in the definition of the context-specific time variable t_s. It is either reset to 0 after each switch,

$$t_s = \infty, 0 \ldots 7, 0 \ldots 7, 0 \ldots 7, 0 \ldots 5, 0, \text{ (switch cost)} \quad (5)$$

or ticks independently in the two contexts,

$$t_s = 0, 0 \ldots 7, 1 \ldots 8, 8 \ldots 15, 9 \ldots 14, 16.$$

$$\text{(independence)} \quad (6)$$

The switch-cost model fits the d' data better than does the independence model: $R^2 = .944$, root-mean-square error of approximation [RMSEA] $= .118$, for Equations 4 and 5 versus $R^2 = .906$, RMSEA $= .154$, for Equations 4 and 6. The number of free parameters (7) is the same. The decisive advantage of the switch-cost model is that it accounts for the recurring costs after the late switches (Blocks 17, 25, and 32), whereas the independence model does not.

A nonparametric analysis corroborates the same conclusion. The independence hypothesis predicts monotonic increases within a given context regardless of any intervening blocks. By our schedule, $d'_{10} > d'_1$, $d'_{18} > d'_9$, $d'_{26} > d'_{17}$, and $d'_{32} > d'_{25}$, where the subscripts stand for block numbers. The three empirical profiles contain 12 such comparisons, and 9 of them turn out negative—against the generally increasing learning trend. The probability of 9 or more such flips happening by chance is 0.073, assuming there is no learning at all. Given the overall positive trend, the probability of this happening under the independence hypothesis is even lower, and hence, it can be rejected. The switch-cost hypothesis, on the other hand, predicts precisely this interference pattern (see Figure 4, top left). The 3 positive d' comparisons in the data occur during the early blocks, exactly as expected.

In light of this evidence, we adopt the switch-cost Equations 4 and 5 as an adequate description of the empirical d' profiles. The best-fitting regression parameters are listed in Table 1. The tem-

poral dynamics of learning appears independent of the difficulty level—the three d' profiles are proportional to one another. Allowing the parameters g, s, τ, and τ_s in Equation 4 to vary freely across the contrast levels i results in virtually no improvement of the fit ($R^2 = .948$, $df = 15$ vs. $R^2 = .944$, $df = 7$), $F(8, 80) < 1$. The three learning curves rise in parallel with a time constant on the order of days ($\tau \approx 10$ blocks = 2.5 days). Superimposed on it are switch costs of constant relative magnitude equal to about 40% of the total d' increase: $s \approx 0.18 \approx 0.40 \, g$. They decay rapidly after the switch that triggers them ($\tau_s \approx 1.2 \ll \tau$) but reappear consistently when the context changes again, for at least 5 switches and 9,600 trials.

Congruence effects. A clear pattern thus emerges in the d' profiles. Does it also emerge when the z-transformed probability correct is plotted separately for congruent and incongruent stimuli? Figure 5 shows that, indeed, the accuracy profiles on incongruent trials mirror the d' profiles: improvement with practice, switch costs, and strong positive effect of the target contrast. The target contrast effect unexpectedly reverses, however, on congruent trials. When the Gabor orientation has the same sign as the context, the accuracy tends to decrease when the target contrast increases! The magnitude of this reversal is small but highly statistically significant, $F(2, 115) = 27.6$, $p < 10^{-10}$.

The inset in the bottom panel in Figure 5 shows the strong interaction between target contrast and congruence, averaged over blocks and observers. The data pattern of each individual participant is virtually identical to the group average. Thus, the counterintuitive reversal for congruent stimuli is a consistent and stable feature of our data.

This reversal is a very powerful constraint on models of perceptual learning. The negative correlation between accuracy and target contrast cannot be explained by any form of response bias, which would trade off the accuracy of congruent and incongruent stimuli but cannot eliminate or alter the ordering of the three contrast levels within the same condition. In fact, a slight bias toward the background orientation is indeed present in the data. The average probability of responding "right" is 0.57 ($z = 0.20$) in the R noise context and 0.43 ($z = -0.15$) in the L noise context. This bias increases the accuracy on congruent trials and decreases it on incongruent ones. However, the response asymmetry by itself does not account for the reversal of the effect of contrast in the congruent condition.

Summary. Overall, the experiment produced a rich data set that reveals an intricate pattern of regularities as summarized in Table 1. Two of these regularities seem especially interesting: the persistent cost for each context switch and the counterintuitive inverse relationship between contrast and accuracy for congruent stimuli. The persistence of the context-specific switch costs in Figure 4 implies that the effects of training in the two stimulus environments are interdependent. This in turn suggests strongly that a common representation and a common decision structure are used in both environments (see Figure 1D). These findings are inconsistent with any model that explains the specificity of perceptual learning solely in terms of nonoverlapping stimulus representations (see Figures 1A–1B) and/or nonoverlapping decision structures (see Figure 1C) because all such schemes entail that the switch cost should diminish with repeated switches and eventually disappear entirely. The switch costs persist undiminished for the full duration of the present experiment—5 switches and 9,600 trials. Coupled with the complex interactions of training, contrast,

Table 1

Summary of the Experimental Results

Basic description	Statistics
Training improves the identification performance in all conditions.	$g = 0.47$ $\tau = 10$
The absolute sensitivity levels depend strongly on the target contrast; the temporal dynamics appears largely independent of it.	$D_{.245} = 2.41$ $D_{.160} = 1.89$ $D_{.106} = 1.20$
Partial context sensitivity: Each reversal of the background context incurs a transient switch cost superimposed on the main learning curve.	$s = 0.18$ $\tau_s = 1.2$
The switch costs persist for at least 5 switches and 9,600 trials.	Figure 4
The identification accuracy of congruent stimuli tends to *decrease* slightly when the target contrast increases.	Figure 5
A small but persistent response asymmetry favors the background orientation.	57% vs. 43%

Note. g = context-general learning; τ = time constant of the context-general learning; $D_{.245}$, $D_{.160}$, and $D_{.106}$ = asymptotic d' levels for target contrasts .245, .160, and .106, respectively; s = switch cost magnitude; τ_s = time constant of the context-specific learning.
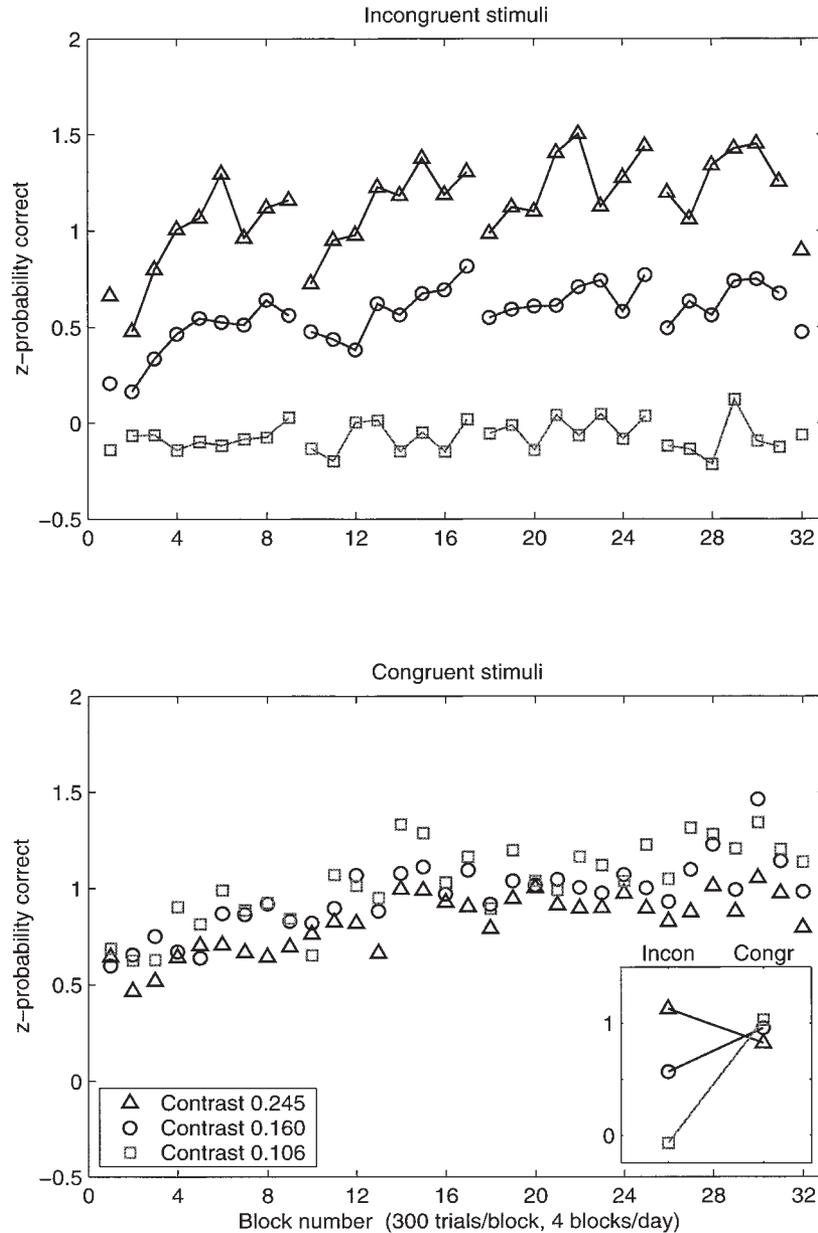
Figure 5. Accuracy profiles for targets oriented against (top) or in the direction (bottom panel) of the background noise (95% confidence interval = ± 0.23). The inset shows the same six curves collapsed in time. Note the paradoxical inverse relationship between contrast and accuracy in the congruent condition.

and congruency, these patterns present a clear theoretical and modeling challenge.

A Reweighting Model of Perceptual Learning

As argued in the introduction, strong convergent evidence suggests that perceptual learning occurs primarily in the connections between early sensory representations and various task-specific response selection areas. Can this *selective reweighting hypothesis* account for the complex data pattern in our experiment? The remainder of the article demonstrates that such account is not only possible but also straightforward and compelling. We present a system that learns by selective reweighting alone—without any

plasticity in the sensory representations—and accounts for all patterns in our challenging data set naturally and quantitatively. Building on the channel-reweighting proposal of Dosher and Lu (1998, 1999), we develop, implement, and test a multichannel reweighting model that takes actual grayscale images as inputs, produces discrimination responses as outputs, and improves its performance on a trial-by-trial basis.

Design Principles

The model consists of two subsystems. The *representation subsystem* encodes the input image as a pattern of activation over a population of units tuned for orientation and spatial frequency

(see Figure 6). This representation scheme is inspired by models of the primary visual cortex. On the basis of this representation, the *task-specific weighting subsystem* determines the discrimination response depending on the activation of a decision unit relative to a response threshold (see Figure 8, which will be discussed later in the text).

The two hypotheses discussed in the introduction have clear-cut interpretation in this simplified framework. *Representation modification* attributes perceptual learning to changes within the representation subsystem; *selective reweighting* (Dosher & Lu, 1998) attributes it to incremental reweighting within the task-specific subsystem.

Both subsystems are intentionally implemented with fairly standard assumptions. The images are processed by a population of units with static tuning properties. There is a single representation layer, a single decision unit, and hence, a single set of modifiable connections (Rosenblatt, 1958). A simple incremental Hebbian learning rule updates the connection weights guided by the local signal-to-noise ratios of the representational units.

Although these properties are inspired by the human brain, the implementation is abstract. The model is not intended to reconstruct the various neural pathways of the visual system and of the

decision-making and action-selection circuits that collectively produce the behavior of interest. Rather, our design strategy has been to reduce the structure of the model to the bare essentials that are still sufficient to perform the task. The intention is to support a specific thesis: Selective reweighting can go a long way toward explaining the detailed patterns of performance in the course of perceptual learning.

The computational model embodies the following five principles of representation and learning.

1. *Orientation- and frequency-tuned representations*: Visual images are represented as patterns of activity over a population of representational units tuned for orientation and spatial frequency. The representational units are meant to be similar in function to simple and complex cells in V1 but may be characteristic of many other visual areas. The model representations ignore other properties of the sensory input, such as color and motion, that are not relevant for the orientation discrimination task.

2. *Contrast gain control*: The representational units are organized in an interconnected network that exhibits some form of feed-forward inhibition, lateral inhibition, shunting inhibition, and/or other nonlinear gain control mechanism(s). In the model, the activation levels of the individual units are dependent on each other and vary as saturating nonlinear functions of the stimulus contrast.

3. *Weighted decision units*: The responses (e.g., "left" or "right") produced on individual trials reflect the activation levels of decision units. Different tasks could be mediated by different pools of decision units. Each decision unit receives inputs, directly or indirectly, from the representational units. The *effective weight* of each representational unit measures the extent to which it influences the activation of the corresponding decision unit.

4. *Incremental associative reweighting*: Perceptual learning in the model occurs through changes in the strength (weights) of the connections (direct or indirect) between the representational units and the decision units (Dosher & Lu, 1998). The learning mechanism is *associative*. When two units—sender and receiver—are active together, the strength of the connection between them increases a little; when one unit is active but the other is not, it decreases a little. The overall performance improves as the system learns to assign higher weights to the stimulus features with high predictive value in the target task and to tune out the distractive and irrelevant features, thereby improving the signal-to-noise ratio at the decision units. This learning is both task- and stimulus-specific. In stationary environments, the weights eventually reach dynamic equilibrium, and the overt performance reaches asymptote. However, the system readjusts its weights whenever the environmental statistics change.
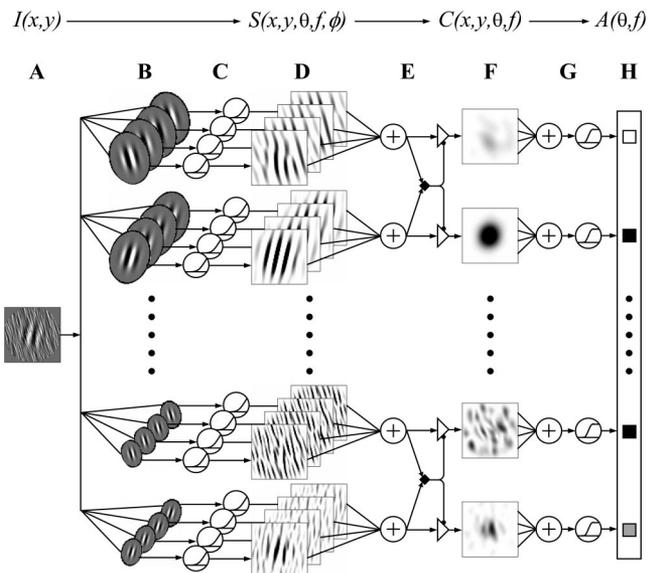

$I(x,y) \longrightarrow S(x,y,\theta,f,\phi) \longrightarrow C(x,y,\theta,f) \longrightarrow A(\theta,f)$

Figure 6. Schematic outline of the perceptual subsystem of the model. Top row: The stimulus $I(x, y)$ is processed by "simple cells" tuned for different orientations θ, spatial frequencies f, and spatial phases ϕ. The resulting *phase-sensitive maps* $S(x, y, \theta, f, \phi)$ are processed further into *phase-invariant maps* $C(x, y, \theta, f)$. Finally, the information is pooled across space by a third population of cells $A(\theta, f)$. The spatial extent of the receptive fields at this stage is commensurate with the diameter of the stimuli ($\approx 2.8°$ of visual angle). A: Stimulus image. B: Schematic rendition of the weight matrices of the various "simple cells." The receptive fields are drawn to scale, and the elliptical aspect ratio is accurate. C: Rectifying nonlinearity (half-squaring). D: Phase-sensitive maps (maximal activations are coded in black). E: Contrast gain control implemented via divisive normalization. This is equivalent to shunting inhibition, depicted by triangles and diamonds. F: Phase-invariant maps. G: Spatial pooling followed by a saturating nonlinearity. H: Final representation $A(\theta, f)$. See the main text and Appendix A for details.

5. *Intrinsic variability*: The processing units throughout the system are noisy, reflecting inefficiencies in receptor transduction and synaptic transmission. This introduces a stochastic component in the internal activation levels, even for a constant input image. Intrinsic variability, combined with the intertrial stimulus variability, limits the accuracy of the overall performance.

Representation Subsystem

The representation subsystem implements orientation- and frequency-selective units and response normalization as outlined schematically in Figure 6. The stimulus image $I(x, y)$ is processed by units tuned to different orientations θ, spatial frequencies f, and spatial phases ϕ. The resulting retinotopic *phase-sensitive maps* $S(x, y, \theta, f, \phi)$ are then processed into *phase-invariant maps* $C(x, y, \theta, f)$. The activation levels throughout these maps are non-negative and normalized (Heeger, 1992). This representational framework is consistent with the principles of both the functional organization of the early cortical visual areas (De Valois & De Valois, 1988; Ferster & Miller, 2000) and the perceptual template model (Lu & Dosher, 1998).

The implementation of the representational subsystem aims at a compromise between parsimony and neurologic plausibility. A MATLAB matrix encodes 35 representational units $A(\theta, f)$ spanning 7 orientations and 5 frequencies. The preferred orientations θ are $0°$, $\pm15°$, $\pm30°$, and $\pm45°$; the preferred spatial frequencies f are 1.0, 1.4, 2.0, 2.8, and 4.0 cycles per degree. Note that the exact orientation of our target stimuli ($\theta = \pm 10°$) is not explicitly included and must be represented in a distributed manner.

The phase-sensitive units $S(x, y, \theta, f, \phi)$ act as linear filters on the stimulus image. Their tuning properties are determined by their input weight matrices (or *receptive fields*), which are modeled by two-dimensional Gabor functions (Adelson & Bergen, 1985; Marčelja, 1980). The orientation bandwidth parameter is $h_\theta = 30°$, a typical value for parafoveal *simple cells* in macaque striate cortex (De Valois, Yund, & Hepler, 1982). The half-amplitude full-bandwidth of the frequency tuning curves is $h_f = 1$ octave, also a typical value. Sensitivity analyses with the model show that these choices of bandwidth, while typical of physiological reports, are not critical to its predictions.

The phase-sensitive maps are then combined into phase-invariant maps (typical of *complex cells*; De Valois et al., 1982; Movshon, Thompson, & Tolhurst, 1978) through a system of *quadrature pairs* similar to those used in energy models of texture and motion perception (Adelson & Bergen, 1985; Knutsson & Granlund, 1983; Pollen & Ronner, 1981). The phase-invariant responses are subjected to nonlinear (divisive) normalization approximating the shunting inhibition in visual cortex (Heeger, 1992). Nonlinearity is a key property both of neuronal responses and of observer models (Carandini, Heeger, & Movshon, 1997; Lu & Dosher, 1999) and is consistent with various interaction effects such as surround suppression and cross-orientation inhibition (e.g., Blakemore & Tobin, 1972; Graham & Sutter, 2000; Tolhurst & Thompson, 1975). In line with physiological and psychophysical evidence, the normalization term in the model is assumed to be essentially independent of orientation and modestly tuned for spatial frequency (Cannon & Fullenkamp, 1991; Carandini et al., 1997; Chubb, Sperling, & Solomon, 1989; Heeger, 1992).

Finally, spatial averaging over the target stimulus region reduces the retinotopic phase-invariant maps to 35 representational units (column H in Figure 6). Like all neuron-like elements, the representation units have limited dynamic range, and their responses saturate for high inputs. Appendix A provides further details about the representation subsystem.

All processing described so far amounts to a deterministic transformation of the input image. The visual system, however, is not deterministic. The variability that is uncorrelated with the stimulus is commonly modeled as *equivalent internal noise* (Ahumada & Watson, 1985; Burgess et al., 1981; Lu & Dosher, 1999). Two such terms are included in the present model: *representation noise* in the representation subsystem and *decision noise* in the task-specific subsystem. Combined with the intertrial variability due to the external noise in the stimuli, these internal noise sources limit the accuracy of the overall performance. In the present experiment, the variability in the representations is dominated by external rather than internal noise.

Figure 7 illustrates the resulting representations. The left panels represent a congruent stimulus (target $\theta = -10°$, noise $\phi = -15°$), and the right panels represent an incongruent one ($\theta = +10°$, $\phi = -15°$). The representations are distributed and noisy, and therefore, the difference between them is rather subtle. Only a few of the 35 units have discriminative value—those tuned to the frequency band of the Gabor target ($f = 2.0$ cyc/deg). Learning to discriminate these representations is accomplished by the task-specific subsystem.

Task-Specific Reweighting Subsystem

The task-specific subsystem carries out the binary classification task. In the model, it is deliberately stripped down to the single-layer network outlined in Figure 8. This simple structure suffices to categorize the stimuli as either "left" or "right" and to improve accuracy with practice. It accomplishes three interrelated functions: information integration, decision, and learning. The model instantiates a strong version of the selective reweighting hypothesis—all learning takes place exclusively through dynamic reweighting of the connections from the representational units to the decision unit. Despite the obvious simplification, the network embodies important computational principles of the brain—parallelism, classification based on weighted feature combinations, and incremental associative learning. The adequacy of the single-layer representation for the present task is considered in the General Discussion.

For optimal discrimination, the system must assign higher weights to those representational units that reliably predict the desired response and lower weights to unpredictive or unreliable units. This kind of problem is ideally suited for implicit statistical learning. Such learning can be accomplished in various ways, and several modeling frameworks are discussed in the literature. These include instance-based models (e.g., Kruschke, 1992; Logan, 1988; Medin & Shaffer, 1978; Nosofsky, 1986), Bayesian classifiers (e.g., Anderson, 1991), decision-bound classifiers (e.g., Ashby, 1992), and connectionist networks (e.g., Grossberg, 1988; Rumelhart, Hinton, & McClelland, 1986). We chose a connectionist network for the present implementation. Although some important differences clearly exist between these alternative approaches, we believe that the principles are very general and could plausibly be instantiated
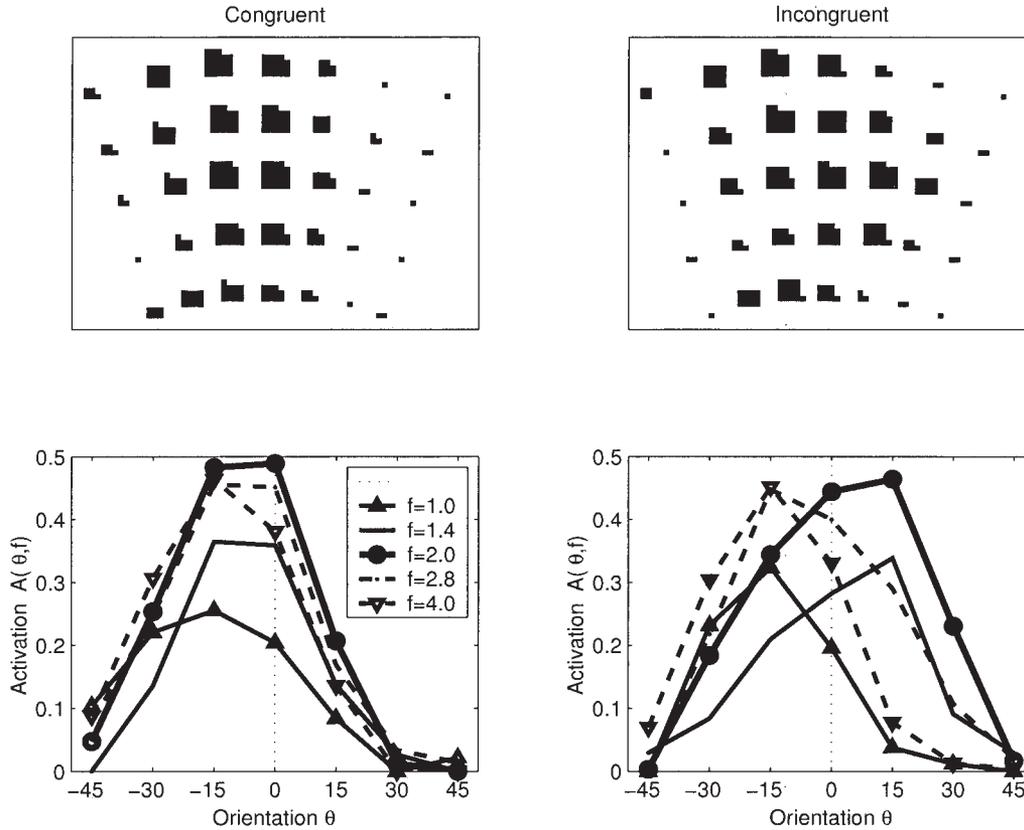
Figure 7. Sample representations of a congruent and an incongruent stimulus in context *L* (negative background orientation). See Figure 2 (top) for the stimuli themselves. Top panels: Hinton diagrams of the 35 activation levels—7 orientations (rays) by 5 spatial frequencies (arcs). The location of the Gabor target is marked by a cross-hair-like symbol in each case. Bottom panels: Line plots of the same representations. Compare with the Fourier spectra in Figure 3. $A(\theta, f) = $ activation of the unit tuned for orientation θ and spatial frequency f.

in other frameworks. We are primarily interested in the learning dynamics on the time scale of hundreds of trials, as revealed by the identification probabilities.[1]

A single decision unit encodes both responses, with negative
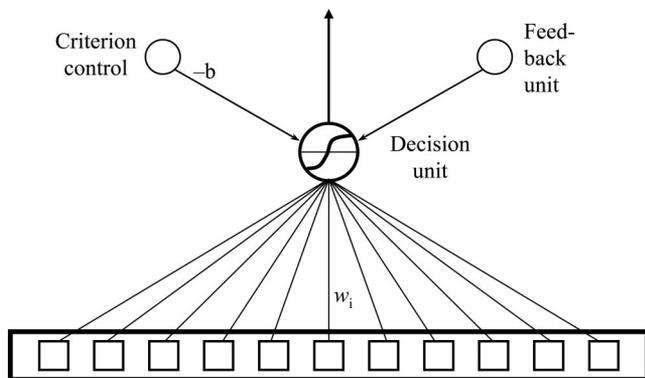

Figure 8. Schematic outline of the task-specific subsystem of the model. Positive inputs to the single decision unit are interpreted as evidence for "right"; negative inputs are interpreted as evidence for "left." The overall sign determines the overt response. Hebbian learning adjusts the read-out connections w_i. $b = $ bias; $w_i = $ read-out connection weights.

activations conventionally signifying "left" and positive ones "right." The overt response is determined by thresholding the decision unit in Figure 8. The stimulus is encoded in the activation pattern a_i across the representational layer (Equation 7; see also Equation 25 in Appendix A). The task-specific subsystem aggregates the bottom-up information using the current weights w_i and adds top-down *bias* b (Equation 8). Gaussian *decision noise* ε with mean 0 and standard deviation σ_d is introduced as well:

$$a_i = A(\theta_i, f_i) \geq 0 \qquad (7)$$

$$u = \sum_{i=1}^{35} w_i a_i - b + \varepsilon = \mathbf{w} \cdot \mathbf{a} - b + \varepsilon. \qquad (8)$$

In connectionist terms, the noisy sum u in Equation 8 is the *net input* to the decision unit. The *output activation o* is a sigmoidal function of u defined by Equation 9.

---

[1] Alternative network modes, such as the leaky, competing accumulator model (Usher & McClelland, 2001), could generalize this system to accommodate any number of response alternatives as well as the dynamics and stochasticity of the decision-making process, including the response-time distributions.

$$o = \frac{1 - e^{-\gamma u}}{1 + e^{-\gamma u}} A_{max} \qquad (9)$$

$$R = \begin{cases} -1 = \text{``Left''} & \text{if } o < 0 \\ +1 = \text{``Right''} & \text{if } o \geq 0 \end{cases} \qquad (10)$$

The final response R is determined by the sign of the output activation o (Equation 10), which in turn equals the sign of the net input u because of the symmetry in Equation 9. Thus, one can calculate exact classification probabilities from the linear Equation 8. The expression $\mathbf{w} \cdot \mathbf{a} - b = 0$ defines a unique hyperplane that bisects the 35-dimensional representation space into "left" and "right" response regions. The hyperplane is orthogonal to the weight vector $\mathbf{w}$ and lies $b/\|\mathbf{w}\|$ units from the origin, where $\|\mathbf{w}\|$ is the euclidean norm of $\mathbf{w}$. Due to the decision noise ε, the response probability varies as a Gaussian function of the distance to this decision boundary (Equation 11). Similar linear classification schemes are used in a wide range of models and theories.

$$P(\text{``Right''}|\mathbf{a}, \mathbf{w}, b) = P(o \geq 0|\mathbf{a}, \mathbf{w}, b) = P(u \geq 0|\mathbf{a}, \mathbf{w}, b)$$

$$= P(\varepsilon \leq \mathbf{w} \cdot \mathbf{a} - b) = \Phi\left(\frac{\mathbf{w} \cdot \mathbf{a} - b}{\sigma_d}\right). \qquad (11)$$

Prior knowledge of orientations is introduced into the model by the initial weight vector $\mathbf{w}(0)$, capturing the empirical fact that the accuracy is well above chance even on Trial 1. The weights are initialized in proportion to the preferred orientation of the units: $w_i = (\theta_i/30°)w_{init}$. The model performance is virtually identical under stepwise initialization, $w_i = sgn[\theta_i]w_{init}$. Values around $w_{init} = 0.17$ produce good fits to the data. The initial weights do not single out the spatial frequency or the specific orientation of the target stimulus; these are factors that must be learned.

Reweighting Via Incremental Hebbian Learning

The model learns via incremental updates of the read-out connections from the representational to the decision unit, thereby instantiating the general principle of incremental Hebbian associative learning. As the experimental data were collected with feedback, it is natural to use a supervised learning rule.

The simulations mirror the experimental procedure and have the correct response available at the end of each trial, represented by the *feedback unit* in Figure 8. During the *early, stimulus-driven phase* the activation of the decision unit encodes the output o that determines the model's response. During the *late, learning phase* the feedback unit applies strong additional input to the decision unit, driving its activation toward $F = \pm A_{max} = \pm 0.5$ for "left" or "right," respectively. The weights are changed at the end of each trial and hence reflect the late, feedback-driven phase:

$$\delta_i = \eta a_i F \qquad (12)$$

$$\Delta w_i = (w_i - w_{min})[\delta_i]_- + (w_{max} - w_i)[\delta_i]_+. \qquad (13)$$

Under the *supervised Hebbian* rule in Equation 12, the change of connection strength i is determined by the product of the presynaptic activation a_i and the postsynaptic activation F. Thus, the task-dependent classification determines the direction, whereas the stimulus features determine the sizes of the individual updates. The model uses *soft weight bounding* (O'Reilly & Munakata, 2000) to keep the weights within bounds. Equation 13 modulates

the basic Hebbian term δ_i by the distance between the current value of the weight w_i and the maximal/minimal value in the corresponding direction.[2] Learning rates η around 0.0015 are consistent with the empirical learning curves.

Adaptive Criterion Control

The activation of the response unit, and hence the accuracy of the response, depend not just on the bottom-up input $\mathbf{w} \cdot \mathbf{a}$ but also on the top-down bias (or decision criterion) b. This bias is necessary to compensate for the influence of the filtered-noise background in the image; it must be adjusted dynamically to track the nonstationary stimulus environment. We assume that the observers try to respond "left" and "right" with approximately equal frequencies, matching the presentation probabilities in the experiment. The bias $b(t+1)$ on each successive trial is set in proportion to the current *response running average* $r(t)$; the distant past is discounted exponentially. The rate parameter ρ is set a priori to 0.02, yielding a time constant of 50 trials. The strength parameter $\alpha = 2.2$ adjusts the effectiveness of the top-down bias relative to the bottom-up inputs in Equation 8.

$$r(t + 1) = \rho R(t) + (1 - \rho)r(t) \qquad (14)$$

$$b(t + 1) = \alpha r(t). \qquad (15)$$

The running average $r(t)$ encodes the current context direction. Each successive change of context triggers a compensatory change of response bias. As b enters Equation 8 with a negative sign, the running averaging introduces a negative feedback loop that stabilizes the model dynamics and promotes balanced response distributions. The feedback loop is reactive rather than proactive, however, and thus cannot fully compensate for the noise context. A small residual preference toward the predominant background orientation persists, as it does in the human responses (cf. Table 1).

Exploring the Perceptual Learning Model

The representational and task-specific subsystems are implemented in a fully functional MATLAB program that takes gray-scale images as inputs and produces classification responses as outputs.[3] The model performance is thus directly comparable to the human data and is tested in a simulation experiment that replicates the psychophysical experiment. The model reproduces the empirical learning curves and accounts naturally for the switch costs and the other phenomena in Table 1.

Simulation Experiment Method

Stimuli and presentation schedule. The stimuli are generated by the same MATLAB software that administered the behavioral experiment. Each image is a Gabor patch embedded in a circular *context* of filtered noise (see Figure 2 and Equation 1). Each training sequence alternates *left* and *right* contexts for a total of 9,600 trials with feedback. There are 1,000

---

[2] The $[x]_+$ operator returns x if $x > 0$ and 0 otherwise; $[x]_-$ returns x if $x > 0$ and 0 otherwise. Thus, repeated potentiation ($\delta_i > 0$) drives the weight up, exponentially approaching the upper bound w_{max}; repeated depression ($\delta_i > 0$) drives it down toward the lower bound w_{mix}.

[3] All data and software are available from the authors and are also available online at http://www.socsci.uci.edu/~apetrov/

runs under schedule *L-8R-8L-8R-6L-R* and another 1,000 under schedule *R-8L-8R-8L-6R-L.*

Procedure. To reduce computer-intensive perceptual processing, a large sample of stimuli was processed in advance to produce a representation pool of 60,000 activation patterns, 2 of which are depicted in Figure 7. They were then sampled without replacement on each run. Fresh internal noise is generated on each run. Each run begins with linear initial weights and no decision bias ($b = 0$). On each trial, the model multiplies the noisy stimulus representation **a** by the current weight vector **w**, adds decision noise and bias, and classifies the stimulus according to the output activation (Equations 7–10). Then it updates the weights slightly and moves on to the next trial (Equations 12–15).

Model parameters and optimization procedure. The model parameters are listed in Table 2. Some of them are set a priori and establish an internal activation scale. The bandwidth parameters of the representation subsystem are chosen on the basis of physiological data detailed in Appendix A. Sensitivity analyses show that none of these choices is critical to the model predictions.

The last six parameters in Table 2 were selected to minimize the summed squared error to the empirical learning curves as defined by Equation 16. The vector *d'* denotes the 96 data points in Figure 4, and $\mathbf{z}_c$ and $\mathbf{z}_i$ denote the corresponding *z*-probability curves for congruent and incongruent stimuli in Figure 5, respectively. Note that $\boldsymbol{d}' = \mathbf{z}_c + \mathbf{z}_i$. The model predictions $\hat{\boldsymbol{d}}'$, $\hat{\mathbf{z}}_c$, and $\hat{\mathbf{z}}_i$ were derived from Equation 11. A general-purpose iterative algorithm (*fmincon*) searched the parameter space.

$$SSE = 2\|\mathbf{d}' - \hat{\mathbf{d}}'\|^2 + \|\mathbf{z}_c - \hat{\mathbf{z}}_c\|^2 + \|\mathbf{z}_i - \hat{\mathbf{z}}_i\|^2. \quad (16)$$

Finally, the model generated 2,000 response sequences with the (approximately) optimal parameters in Table 2. The binary responses were tabulated, counted, and analyzed in exactly the same way as were the behavioral data, by the same software.

Simulation Results and Discussion

There are three independent variables: block, difficulty, and congruence. The context manipulation is embedded within the time variable block.

Congruence effects. One of the most surprising and theoretically challenging aspects of the empirical data is the effect of the congruence between the Gabor orientation and the predominant background orientation. Figure 9 plots the model predictions for incongruent and congruent stimuli averaged across the 2,000

Table 2
Parameters Used in the Simulation Experiment

Parameter	Value
Parameters set a priori	
Orientation spacing	$\Delta\theta = 15°$
Spatial frequency spacing	$\Delta f = 0.5$ oct
Maximum activation level (Equations 9 and 26)	$A_{\max} = 0.5$
Weight bounds (Equation 13)	$w_{\min/\max} = \pm 1$
Running-average rate (Equation 14)	$\rho = 0.02$
Parameters constrained by published data	
Orientation tuning bandwidth	$h_\theta = 30°$
Frequency tuning bandwidth	$h_f = 1.0$ oct
Radial kernel width (Equation 24)	$h_r = 2.0$ dva
Parameters optimized to fit the present data	
Decision noise (Equation 8)	$\sigma_d = 0.195$
Representation noise (Equation 24)	$\sigma_r = 0.100$
Activation function gain (Equations 9 and 26)	$\gamma = 0.80$
Learning rate (Equation 12)	$\eta = 0.0015$
Initial weight scaling factor	$w_{init} = 0.17$
Bias strength (Equation 15)	$\alpha = 2.2$

Note. Six parameters are fitted to the data; the rest are set independently.

model runs. All qualitative patterns in the data reappear in the model predictions. In particular, the three incongruent curves (top panel) are widely spaced, whereas the congruent ones are almost on top of each other (bottom panel), exactly as their empirical counterparts in Figure 5. The model also accounts for the counterintuitive slight reversal of the target contrast effect for congruent stimuli. Just as with the human observers, it is slightly less accurate for high-contrast congruent targets than for low-contrast ones (see Table 3 and the inset of Figure 9).

Learning curves. Figure 10 plots the *d'* learning curves averaged across the 2,000 model runs. All qualitative patterns in the psychophysical data are clearly replicated: the slow but steady general improvement, the parallel dynamics across the three target contrast levels, and the recurring switch costs. Figure 10 tracks five context switches, and additional model simulations confirm that the pattern persists indefinitely, long after the context-general component levels off. The cost magnitude does not diminish with successive switching.

Quantitatively, the model accounts for 88% of the variance of the 96 data points in Figure 10 ($R^2 = .882$, RMSEA $= .209$). This approaches the fit of the regression Equation 4 discussed earlier ($R^2 = .944$, RMSEA $= .118$) and is very good performance given that the regression merely describes the data whereas the model provides a principled, mechanistic explanation. In particular, the clear separation between the three target contrast levels in the model *d'* is driven entirely by the differential signal-to-noise ratio of the input images. Unlike Equation 4, the model has no adjustable parameters controlling the vertical placement of the individual curves. The good parameter-free approximation of this property of the data testifies to the general appropriateness of the representation subsystem. Fine tuning of the representation nonlinearity[4] probably could improve the fit even further, but it is already clear from Figure 10 that the model meets its objective to provide an existence proof of the empirical adequacy of the reweighting hypothesis.

Weight dynamics. The evolution of the weights under the Hebbian learning rule is well worth examining in detail. Figure 11 shows it in diachronic (top row) and synchronic cross-sections (middle and bottom rows). Early in training (panel "*T* = 0300"), the weights are still close to their initial pretraining values and are thus linearly related to the corresponding orientation regardless of spatial frequency. As training progresses, the selective reweighting mechanism strengthens some connections and weakens others, gradually adapting the system to the statistical structure of the stimulus environment.

The top right-hand panel in Figure 11 traces the strength of the units tuned to *f* = 4.0 cyc/deg. The seven lines correspond to the various preferred orientations of the representational units, ordered from $\theta = -45°$ (bottom line) to $45°$ (top line). The weights of all these units grow weaker over time, reflecting the fact that they are tuned to a spatial frequency band containing a lot of visual noise but virtually no signal. As their activation does not correlate with the output activation, the net Hebbian driving signal δ in Equation 12 is close to zero and the weights gradually decrease via the

---

[4] The present model uses simple squaring (power 2) to rectify the phase-invariant maps (see Figure 6C). Estimates of transduction nonlinearities for the perceptual template model from multiple-criterion threshold data (see Lu & Dosher, 2000) over many data sets have yielded powers in the range of 1.7 to 2.5.
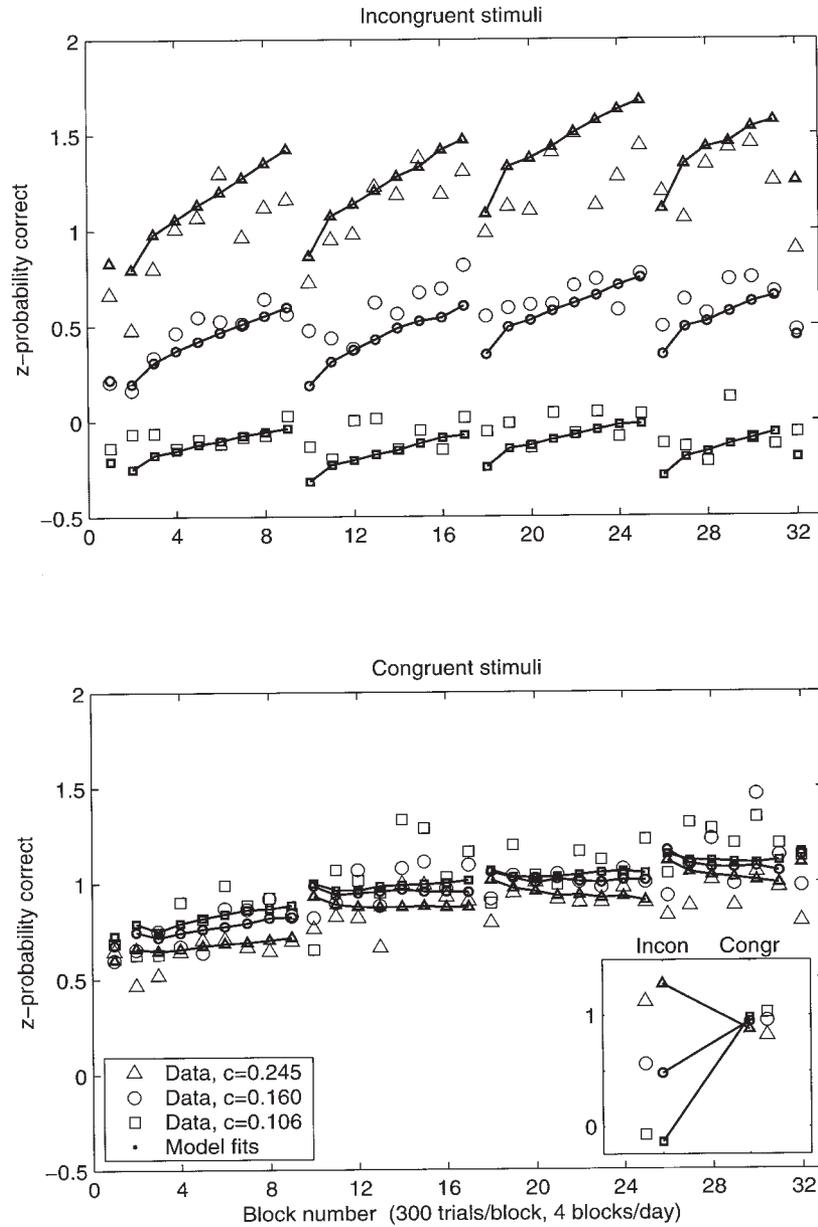
Figure 9. Accuracy profiles for targets oriented against (top panel) or in the direction (bottom panel) of the background noise. The inset shows the same six curves collapsed in time. Compare with Figure 5.

nonlinear bounding scheme in Equation 13. The system thus "tunes out" irrelevant frequency bands, in agreement with the experimental literature (Dosher & Lu, 1998; Saarinen & Levi, 1995).

At the same time, the system "tunes into" the relevant spatial frequencies and orientations. The top left-hand panel in Figure 11 shows the weights of the units tuned to the target frequency $f = 2.0$ cyc/deg. Of these, the units tuned to orientations $\theta = \pm 15$ and ± 30 are strengthened because they are activated by the targets whereas the ones tuned to $\theta = \pm 45$ are not. (The peak spectral power of the Gabor targets is centered on $\theta^* = \pm 10$; see Figure 3.) The midline unit ($\theta = 0$) is activated equally by either stimulus, and thus, its weight remains zero. This reflects the

associative nature of the learning rule: The unit is not strengthened despite its high activation on every trial because this activity does not contribute to the task. Consistent exposure is a necessary but not sufficient condition for perceptual learning.

Moreover, the units most responsive to the two target orientations are not the ones that develop the strongest weights. The strongest weights in Figure 11 are tuned to $\theta = \pm 30$, although the units that best match the target stimuli are those tuned to $\theta = \pm 15$. The most diagnostic units for orientation discrimination are not the most active ones but those with the highest signal-to-noise ratio. It is interesting that the only positive report of training-induced representation modification in V1 known to us also involves precisely the "neurons with preferred orientations lying between

Table 3
Accuracy (Average z-Probability Across Blocks) as a Function of Target Contrast and Congruence (See Figures 5 and 9)

	Peak target contrast c_p			
Stimulus	0.106	0.160	0.245	Total
Empirical data ($CI_{95} = \pm 0.040$ within blocks)				
Congruent	1.031	0.956	0.821	0.936
Incongruent	−0.070	0.565	1.126	0.541
Total ($= d'/2$)	0.481	0.761	0.974	0.738
Simulation results				
Congruent	0.981	0.947	0.874	0.934
Incongruent	−0.138	0.482	1.284	0.542
Total ($= d'/2$)	0.422	0.714	1.079	0.738

Note. CI_{95} = 95% confidence interval; c_p = peak target contrast.

12 and 20° of the trained orientation" (Schoups et al., 2001, p. 551). Our Hebbian rule discovers the diagnostic value of these units from the statistics of the learning corpus. Had the task been detection rather than discrimination, the weights would arrange themselves differently.[5] Koyama, Harner, and Watanabe (2004) found the same kind of repulsion in perceptual learning of motion-direction discrimination.

There are some subtle but very important asymmetries in the trained weight vectors in Figure 11. In context *R*, the negative ("left") orientations develop stronger weights than do the positive ("right") ones. This happens because the background noise injects irrelevant variability into the *R* "channels" and drives their weights down. The pattern reverses in context *L*. The model effectively looks for incongruent targets standing out of the noise. This context-dependent weight asymmetry leads to switch costs when the context changes. Modest suboptimality of a few critical weights suffices to induce a noticeable decrease (about 15% in Figure 10) of the system-wide discriminability after each switch.

Sensitivity analysis. The ability of the model to account for the qualitative patterns in the data does not appear overly dependent on the particular parameter settings in Table 2. We repeated the simulations many times with various parameters set above or below their optimal values. The results show the same qualitative pattern in all cases and often match the ones reported here quantitatively as well. The switch cost is very robust, though sometimes small. The contrast reversal for congruent stimuli is the most dependent on specific parameter ranges.

Formal Analysis of the Model

The simulation experiment demonstrated the model's ability to reproduce the human behavioral patterns. Auxiliary simulations demonstrated that this ability is not restricted to a narrow parameter range. Finally, mathematical analysis reveals the statistical foundations of this robust behavior.

Statistical properties of the stimulus environment. The perceptual subsystem maps each image to an activation vector in the 35-dimensional representation space, thereby transforming distributions of external stimuli into distributions of internal representations. Figure 12 illustrates these internal distributions in various projections. There are 12 clusters corresponding to the 12 stimulus classes. The substantial variability within each cluster is due to external image noise and internal perceptual noise. Coupled with

the internal decision noise, this variability limits the classification accuracy the system can achieve. The main objective of the learning mechanisms is to adjust the weight vector **w** and the decision criterion *b* to maximize the accuracy under these constraints. The optimal solution depends on the means, variances, and covariances of the representations.

We calculated the covariance matrices of the 12 representation pools from the simulation experiment and subjected them to a hierarchical cluster analysis. The 6 matrices within each context are very similar to each other (point-wise correlation $r = .87$, on average) and different from the 6 matrices in the other context ($r = .30$). The elliptical equal-probability contours at the top of Figure 12 illustrate they have nearly identical orientation and shape within each context and very different orientations between contexts (thin lines = *L*, bold lines = *R*). Thus the background noise determines the covariance structure of the representation distributions, whereas the Gabor targets determine mostly their centroids in the 35-dimensional space. This property of the representations reflects the spectral distributions of the images (see Figure 3) and stems from the highly variable nature of the background noise and the deterministic nature of the Gabors.

As the context is fixed in any given experimental block, the covariance matrices of the 6 distributions the system is confronted with are approximately equal.[6] This equality has far-reaching theoretical implications. First, when two (multivariate normal) distributions have equal covariance matrices, the optimal decision bound between them is linear (Ashby & Gott, 1988). This in turn suggests that the simple one-layer network in Figure 8 can attain near-optimal performance on the present task, justifying its adoption in the model.

The second implication of the approximately equal covariance structure is that the discriminability d' does not depend on the decision criterion *b*. For every pairwise classification, d' depends only on the weight vector **w**. (See Equation 30 in Appendix B for the specific formula.) The weights can thus be learned independently of the criterion, which greatly simplifies the learning of each. Also, as long as each target contrast level is analyzed individually, the d' predictions (e.g., see Figure 10) are not contingent on any particular criterion control strategy.

Global optimality through local mechanisms. The weight vector that maximally separates two given stimulus classes can be derived from the corresponding population statistics through linear algebra (e.g., Ashby, 1992). To a good approximation discussed in Appendix B, the optimal weight of each unit is proportional to its local signal-to-noise ratio:

$$w_i^* \propto \frac{\langle a_i \rangle_R - \langle a_i \rangle_L}{\sigma_{a_i}^2 + c^2}. \tag{17}$$

In Equation 17, $\langle a_i \rangle_R$ is the average activation of unit *i* across the stimuli classified as "right," $\langle a_i \rangle_L$ is the analogous average for

---

[5] A similar pattern has been observed experimentally in adaptation (Regan & Beverley, 1985). Adapting to a high-contrast grating degraded discrimination for test gratings inclined at about 10°–20° to the adapting grating while having little effect on the detection of these inclined gratings. For test grating parallel to the adapted grating, discrimination was improved, but detection was degraded.

[6] Also, the stimulus-independent decision noise in Equation 8 attenuates the slight existing inequalities. See Appendix B for details.
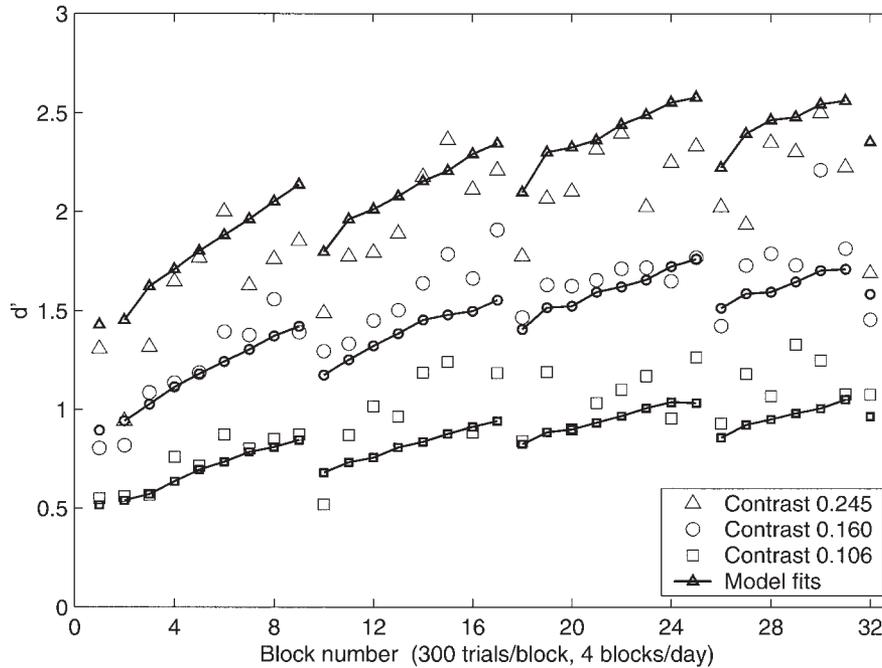
Figure 10. The learning curves generated by the model (solid lines) reproduce the empirical *d'* values (open symbols) well. Compare with Figure 4.

"left" stimuli, $\sigma_{a_i}^2$ is the variance of a_i within either population, and $c^2 = \sigma_d^2/\|\mathbf{w}\|^2$ is a constant fraction of the decision noise in Equation 11. All stimuli have the same background and the same target contrast. The overall learning task involves a mixture of three such pairwise comparisons. This is not a serious problem because the corresponding optimal vectors are very nearly parallel to each other, and hence, a common decision hyperplane maximizes the discriminability across all three target contrasts as long as the embedding context remains stationary.

Equation 17 has two very important properties: It is *local* and *associative*. The weight of a representational unit does not depend on the activity of the other representational units. It does depend, however, on the activation of the decision unit because the latter encodes the task-dependent category label *L* or *R* (cf. Figure 8). That is why the connection between these two units is the true locus of learning. The presynaptic activation a_i carries information about the stimulus; the postsynaptic activation during the late phase of the trial carries information about the desired response. This information converges to the synapse where it is integrated over time, extracting the population statistics from experience. This is all that is needed to achieve near-optimal accuracy of the system as a whole, in precise agreement with the selective reweighting hypothesis (Dosher & Lu, 1998).

The supervised, bounded Hebbian learning rule in Equations 12 and 13 sets the weights remarkably close to their optimal values in the long run because it tracks the statistics prescribed by Equation 17. Specifically, when averaged over many trials with a slow learning rate η, the Hebbian term δ_i in Equation 12 is proportional to the numerator in Equation 17. As the postsynaptic activation *F* equals $+A_{max}$ for "right" and $-A_{max}$ for "left" responses,[7] it serves as an indicator variable for the two categories:

$$\langle \delta_i \rangle = \eta \langle a_i F \rangle = \eta A_{max}[\langle a_i \rangle_R - \langle a_i \rangle_L]. \tag{18}$$

The associative learning rule thus drives each weight in proportion to the difference between the conditional averages $\langle a_i \rangle_R$ and $\langle a_i \rangle_L$. The variability of the representations is taken into account as well, even though the explicit division by $\sigma_{a_i}^2 + c^2$ in the ideal Equation 17 is biologically unrealistic. The weight-bounding mechanism approximates the desired effect indirectly by penalizing the highly variable units relative to the less variable units. For instance, assume a weight $w > 0$ is updated in opposite directions δ and $-\delta$ on consecutive trials. Because of the saturating nonlinearity in Equation 13, the negative update has greater impact than does the positive update, resulting in an overall regression toward the neutrality point $w = 0$. The more variable the updates—for example, $\pm 2\delta$ instead of $\pm \delta$—the stronger the regressive effect. An inverse relationship between connection strength and activation variability emerges, approximating the optimal solution.

Statistical learning in nonstationary environments. The underlying cause of the recurring switch costs in Figures 4 and 10 is now apparent: The two environments have different statistical structure, and therefore, weights optimized in one context are not optimal in the other. Specifically, the background affects strongly the variances $\sigma_{a_i}^2$ in Equation 17—channels that are free of noise in one context are clobbered with noise in the other. The numerator of Equation 17 is also context dependent despite the subtraction operator there because of the nonlinearities in the representation subsystem. Thus, no common set of weights exists that can max-

---

[7] This choice ensures the condition $\langle F \rangle = 0$. A more general and biologically realistic implementation would augment Equation 12 with a sliding baseline: $\delta_i = \eta a_i (F - \langle F \rangle)$. Such sliding thresholds are very common in Hebbian models and appear well grounded physiologically (e.g., Bienenstock, Cooper, & Munro, 1982).
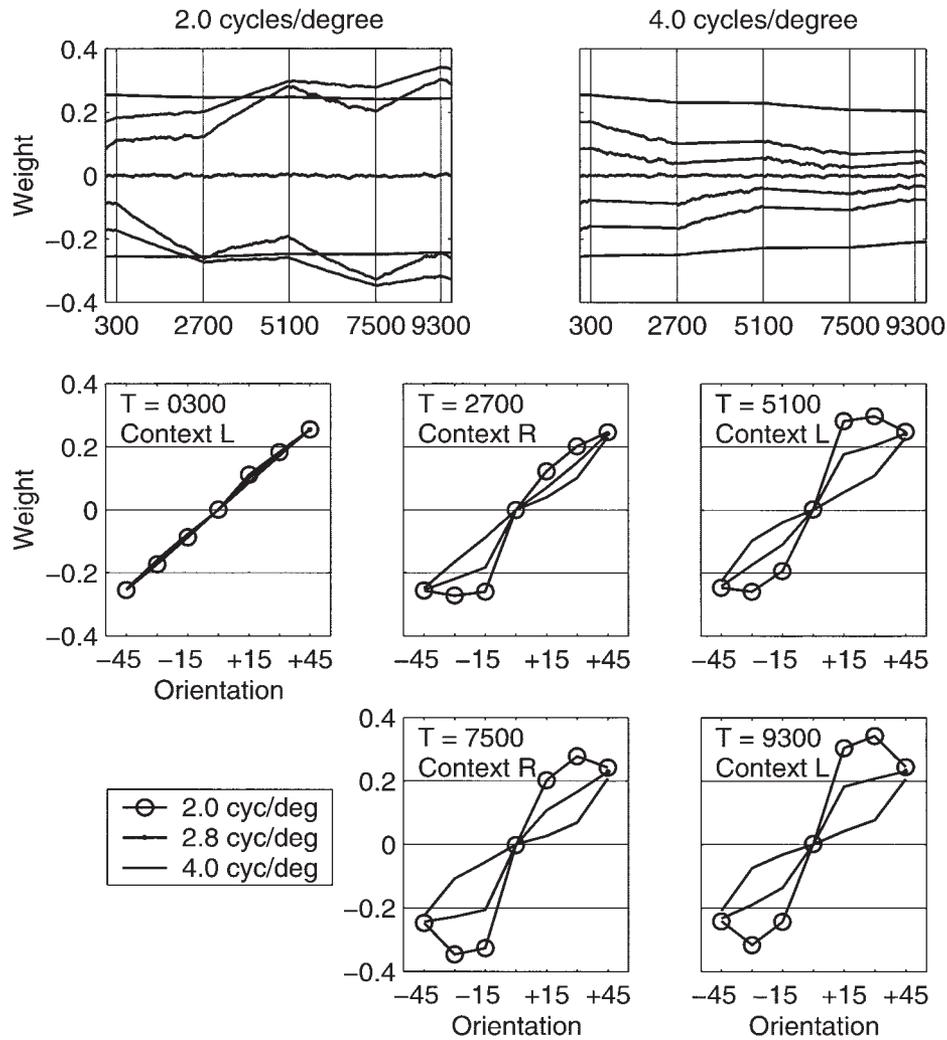
Figure 11. Weight dynamics of a typical model run under schedule L-8R-8L-8R-6L-R. Top row: Longitudinal weight traces for units tuned to the target frequency (2.0 cyc/deg) and an irrelevant frequency (4.0 cyc/deg). Each trace corresponds to a particular orientation. Middle and bottom rows: Cross-sections at the end of each epoch. Not all frequencies are shown. Note that the weight vector differs across the two contexts, hence the switch costs. See text for details. L = left; R = right; T = trial number.

imize the discriminability in both contexts. The right panel in Figure 12 illustrates this deep-seated asymmetry.

The system continuously adapts its weights to the statistics of the environment at hand. As only one set of connections is available, their adjustment in one context disrupts earlier adjustment in the other. When the context switches again, the system lags behind with suboptimal weights until it readapts. The recurring switch-cost pattern in Figures 4 and 10 is a signature manifestation of these repeated adjustments. This pattern strongly suggests that the human observers, like the model, depend on a single decision pathway in this task and learn via statistically driven selective reweighting.

Explanation of the Empirical Phenomena

The empirical phenomena listed in Table 1 follow naturally from the architectural principles advocated in this article and instantiated in the model.

Training improves the identification performance in all conditions. The associative learning mechanism optimizes the system for the task at hand. It assigns higher weights to the units tuned to frequencies and orientations that predict the desired output and lower weights to the irrelevant units. This selective reweighting improves the signal-to-noise ratio of the aggregate decision variable.

The absolute sensitivity levels depend strongly on the target contrast; the temporal dynamics appears largely independent of it. The first phenomenon reflects the intrinsic strength of the stimulus inputs; the second follows directly from the fact that the same set of weights is applied to all stimuli at any given moment in time.

Each reversal of the background context incurs a transient switch cost superimposed on the main learning curve (partial context sensitivity). The two environments have different statistical structure. Consequently, weights optimized in one context are
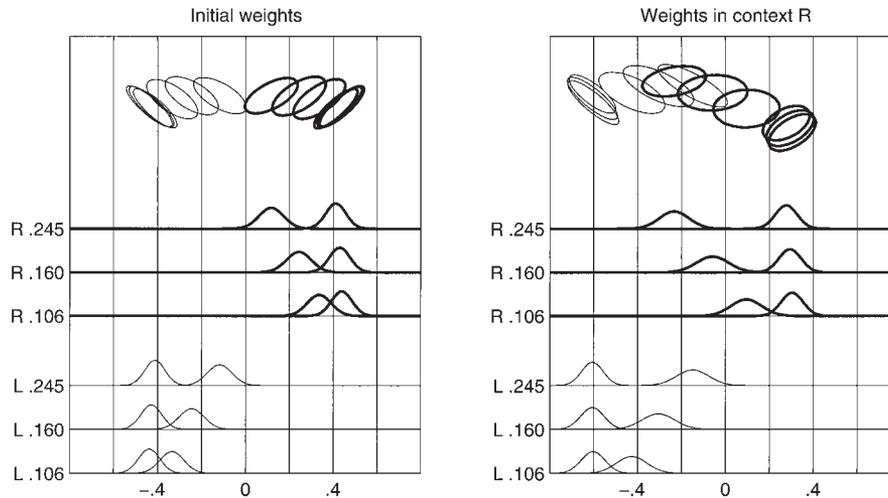
Figure 12. Projections of the 35-dimensional representation space onto 2 dimensions (top) or the single decision dimension (bottom). The ellipses enclose 90% of the mass of the 12 stimulus populations, and the bell curves plot the corresponding probability densities of the weighted combination **w** · **a** separately for the three contrast levels in each context. Left panel: The initial weight vector discriminates the "left" and "right" targets, but the distribution overlap can be reduced. Right panel: After extensive training in context*R*, all stimulus classes have become more discriminable in the projection defined by the new weight vector, especially those in context*R* (solid lines). If the context now changes to L (thin lines), a switch cost will result. L = stimuli in context*L*; R = stimuli in context*R*; **w** · **a** = dot product of the weight vector **w** and the activation vector **a**.

not optimal in the other. For relatively slow learning rates, the system lags behind and temporarily works with suboptimal weights after each switch. Learning is slow in order to accrue the large samples necessary to separate the genuine regularities from the chance fluctuations.

The switch costs persist for at least 5 switches and 9,600 trials. The two contexts activate extensively overlapping representations and depend on the same connections, as illustrated in Figure 1D. Given the structure of the environment, no common set of weights exists that maximizes the discriminability in both contexts.

The identification accuracy of congruent stimuli tends to decrease slightly when the target contrast increases. This effect reflects compressive nonlinearities in the representation subsystem. When a congruent stimulus is presented, some activation "leaks" into the incongruent representational units because of their relatively broad orientation tuning. At the same time, the congruent units are essentially saturated by the background noise and, therefore, are insensitive to the contrast of a congruent target. Thus, the activation of the congruent units is effectively constant, whereas the activation of the incongruent units increases when the target contrast increases. Elevated activation of the incongruent units is interpreted as evidence against the congruent (correct) response, and accuracy and contrast become inversely related.

A small but persistent response asymmetry favors the background orientation. The congruent representational units, and by extension the congruent response, consistently receive activation from the noisy background texture. The criterion control machinery compensates for most of this imbalance, but not all.

General Discussion

In this article, we first provided a novel task analysis framework to organize the prior research on learning and transfer in perceptual

tasks. A particular task type with overlapping representations and decision structures was identified as diagnostic for testing existing claims about the locus of perceptual learning. An experiment motivated by this analysis yielded a highly articulated set of data from the relevant task category. On the basis of known principles of the human perceptual system, we designed and implemented a new multichannel reweighting model of perceptual learning. We evaluated the model both by direct fits to experimental data and by mathematical analysis. A statistically driven, linear classification scheme was sufficient to account for the learning dynamics, switch costs, and several other phenomena in the present task and, we suggest, many other perceptual learning tasks. In this final section, we discuss the multichannel reweighting architecture in light of the available behavioral and physiological data, compare it to related models, and derive some predictions.

Task Framework and Empirical Tests

The stimulus specificity and task specificity of perceptual learning, widely cited as evidence for an early plasticity site, is diagnostic in only a subset of cases. In Figure 1, we outlined a taxonomy of situations that test specificity or, conversely, generalization. Designs that rely on nonoverlapping representation populations and, hence, nonoverlapping sets of read-out connections cannot discriminate between representation modification and task-specific reweighting as possible learning mechanisms. Both hypotheses are equally consistent with the observed specificity in these common designs. Specificity could provide strong evidence for change in the early representations only in carefully controlled situations that engage highly overlapping representations in two task environments. Very few studies have explored such designs. Here, we specify a task that uses the same representational units in each of two stimulus environments, with slightly different patterns

of activation across these units. A nonstationary manipulation of the noisy background texture induced recurring switch costs in the learning curves, verifying that the two tasks share a common set of connections as well.

Representation Modification Versus Reweighting

This same-representations learning experiment provides a challenging test of the task-specific reweighting mechanisms in the absence of any significant alteration of the sensory representations. An extremely parsimonious architecture with incremental Hebbian reweighting over fixed multichannel sensory representations is sufficient to provide an excellent account of this challenging data set.

These results are generally consistent with the current research evaluating the impact of perceptual learning on the orientation and spatial frequency selectivity of cells in early sensory cortices (V1, V2). This research has found very little evidence of substantial modification of these early representations solely as a result of behavioral practice in adults. Two forms of plasticity have been examined in detail: *recruitment* and *sharpening*. The former involves changes in the preferred orientation of the representational units; the latter involves changes in the slope (or, equivalently, bandwidth) of their tuning curves. As described in the introduction, three experiments recorded from single cells in intact striate cortices of adult monkeys following extensive practice (Crist et al., 2001; Ghose et al., 2002; Schoups et al., 2001). None of these studies found any evidence of recruitment, and only one study (Schoups et al., 2001) detected modest sharpening, which was not replicated in a similar experiment (Ghose et al., 2002), although Yang and Maunsell (2004) found some sharpening of tuning curves in V4.

On the other hand, there is considerable physiological evidence of rerepresentation in other modalities as well as after invasive manipulations in vision. Representation modification is therefore a clear possibility. Different mechanisms for such modification have been proposed, including recruitment, sharpening, synchronization, and multiplexing (task-dependent dynamic retuning; Gilbert et al., 2001). The available physiological reports in the visual domain do not support a substantial role for recruitment in perceptual learning (Gilbert et al., 2001). Tuning-curve sharpening was considered in detail as a possible mechanism by Teich and Qian (2003), who showed that if sharpening does occur, it could provide an explanation for certain accuracy improvements, although perhaps not for the full extent of observed improvements (e.g., Yang & Maunsell, 2004). Also, their proposal provides an incomplete explanation because no mechanistic description of how such sharpening occurs incrementally over time is specified. Furthermore, the changing activation patterns resulting from such modification would necessitate additional reweighting to be interpreted correctly by subsequent decision processes.

The synchronization and multiplexing mechanisms of representation modification (e.g., Gilbert et al., 2001; Li et al., 2004), absent physiological specification and testing, may be functionally indistinguishable from a reweighting mechanism working on stable representations. Synchronization may serve to increase the net activation of select representational units, which is functionally equivalent to selective strengthening of the connections linking these units to higher areas. Multiplexing, in which responses of units change only in a particular task environment, may also be

functionally equivalent to reweighting in that the representations are stable. Practice in one task does not alter the representations for another task.

The present study addresses the representation modification hypothesis because the stimuli occupy the same regions of orientation space in both training contexts. Sharpening the orientation selective units may account for some, but not all, of the observed behavioral improvement, as discussed below. However, no form of recruitment or sharpening seems able to account for the paradoxical inversion of contrast effects in incongruent stimuli.

From a functional point of view, stability of the early sensory representations has considerable appeal. In the intact visual cortex, the recruitment of units for one orientation, for example, might occur at the expense of the other orientations. Analogously, reorganization for the benefit of a particular task is likely to interfere with other tasks, given that the same early areas serve them all. Such eager policy could easily jeopardize the integrity of the representations. In the words of Ghose et al. (2002), "It would seem undesirable to change early representations so as to improve performance in a particular task unless there was no alternative" (p. 1884). Our reweighting model proves there is an alternative. Moreover, the horizontal collaterals in V1 are implicated in a range of functions, including orientation sensitivity (Nelson, Toth, Sheth, & Sur, 1994; Somers, Nelson, & Sur, 1995) and contrast gain control (Carandini et al., 1997; Heeger, 1992; Ohzawa, Sclar, & Freeman, 1982). It is not clear whether and how these collaterals could accommodate the additional function of perceptual learning. This may be one important difference between vision and other modalities—The primary visual cortex performs a more complex computation that imposes tighter functional constraints on the neural substrate.

In conclusion, the hypothesis that perceptual learning is due predominantly to representation modification implemented in the lateral connections is not an obligatory implication of existing behavioral or physiological observations. Moreover, the lateral connections are under tight functional constraints, and it remains an open research question as to whether they can accommodate the additional function of perceptual learning. A much more natural and computationally efficient way to improve performance is to adjust the strength of the connections from the early representations to subsequent decision processes, giving high weight to the reliable and predictive features and low weight to the unreliable or irrelevant ones. This selective reweighting hypothesis is consistent with the specificity of perceptual learning. As discussed in the introduction, it accounts for certain experimental results better than does the representation modification hypothesis (Ahissar, Laiwand, Kozminsky, & Hochstein, 1998; Crist et al., 1997; Dosher & Lu, 1998, 1999, 2005; Fine & Jacobs, 2000).

The distinction between these two hypotheses becomes less clear in multilayer networks because changing the weights also changes the activation patterns at the intermediate layers. In principle, learning may occur at multiple levels of the visual processing hierarchy (Ahissar & Hochstein, 1997, 2004). The representation modification hypothesis, however, is typically advanced with respect to the very early areas, particularly V1, where the two hypotheses are distinct (e.g., Karni & Sagi, 1991). Representation modification presumably involves the lateral connections within V1, whereas selective reweighting involves the projections from V1 to other areas. The main thrust of the selective reweighting hypothesis is that perceptual learning occurs predominantly

through incremental associative updating of the projections between different areas, as opposed to updating the lateral connections within a specific area.

How Far Can Representation Modification Go?

Our model assumes fixed representations and learns exclusively by reweighting the read-out connections. The excellent fits to the data demonstrate that the selective reweighting mechanism is sufficient to account for the behavioral phenomena. Can the same phenomena be accounted for by a representation modification mechanism? The natural way to answer this question is to build a model that assumes fixed read-out connections and learns exclusively by modifying the representations. This, however, would be an extraordinary achievement. Processing the input images into biologically realistic activation patterns requires a fairly complex set of processors, as evident from the schematic outline of the perceptual subsystem in Figure 6. A representation modification model would have to design a method for incremental alteration of this processing and, in addition, satisfy several strong behavioral constraints. First and foremost, representation modification should be able to reproduce the full extent of the observed behavioral improvement. Threefold reduction of orientation discrimination thresholds is not uncommon for humans (e.g., Schoups et al., 1995), and even more dramatic reduction is routine for monkeys (e.g., Schoups et al., 2001; Yang & Maunsell, 2004). Second, training on one stimulus at one location should not degrade the discriminability of other stimuli at other locations. Third, the learning effects should last virtually indefinitely without further practice, resisting interference from the constant stream of visual stimulation. Many other constraints can be listed (Tsodyks & Gilbert, 2004), and it appears that even a system as exquisite and intricate as the primate visual cortex cannot satisfy them all.

Representation modification considered in isolation seems insufficient to account for the full extent of perceptual learning. But how far can it go? Consider tuning-curve sharpening as the best documented form of representation modification. Sharpening is the main finding in the two positive reports of training-induced changes in intact early visual cortex of adult monkeys. Concretely, the strongest empirical evidence of representation modification to date amounts to approximately 30% reduction of the orientation tuning bandwidth of select neurons in V1 (Schoups et al., 2001) and approximately 13% reduction in V4, barely reaching statistical significance (Yang & Maunsell, 2004). At the same time, the psychophysical thresholds improved at least tenfold in both studies! This dramatic behavioral effect cannot be accounted for by such modest sharpening even if one assumes a maximally efficient decoding scheme (ideal Bayesian observer; Zhang, Ginzburg, McNaughton, & Sejnowski, 1998, as cited in Schoups et al., 2001).

To evaluate the potential of sharpening to account for our data, we modified the multichannel reweighting model to set the orientation bandwidth parameter h_θ individually for each channel. Presumably, such selective sharpening can be accomplished through subtle changes in the recurrent connectivity in V1 (Teich & Qian, 2003). For completeness, we also allowed individualized modification of the frequency bandwidth h_f, although we are not aware of any empirical or theoretical claims of sharpening along the spatial frequency dimension. Keeping all weights fixed at all times, we measured the overall model performance as a function of systematic changes in the tuning bandwidths.

Figure 13 plots the relative change in the average d' assuming linear weights $w_i \propto \theta_i$ (left panel) or stepwise weights $w_i \propto sgn[\theta_i]$ (right panel). Note that these weight vectors single out neither the orientation ($\theta^* = 10°$) nor the spatial frequency of the Gabor targets. Any improvement in the observer-level d's must result from sharpening the tuning curves of select representational units. One set of simulations varies the orientation tuning bandwidth of the units with preferred orientation between $-30°$ and $30°$ and preferred spatial frequency matching the Gabor frequency ($f = 2.0$
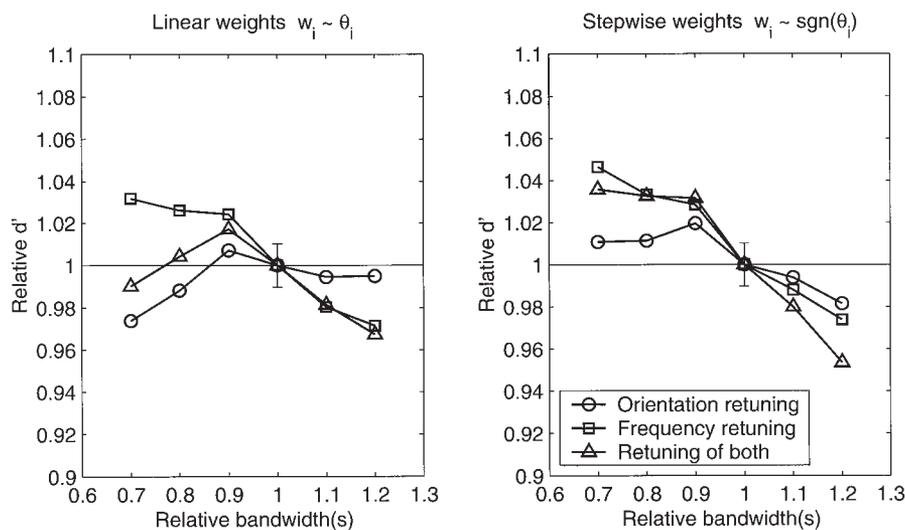

Figure 13. Running the model with fixed weights but sharper tuning curves of select units produces only modest behavioral improvements. The abscissa scales the bandwidth of orientation tuning, spatial frequency tuning, or both, relative to the default bandwidth parameter for the corresponding dimension(s). The error bar denotes 95% confidence interval. w_i = connection wait of unit i; θ_i = preferred orientation of unit i; sgn = sign of.

cyc/deg). The orientation bandwidth h_θ of these five channels is varied systematically between 21° and 36°; the bandwidth of all other channels is fixed to the default $h_\theta = 30°$. This is a reasonable approximation of both the single-cell recordings (Schoups et al., 2001; Yang & Maunsell, 2004) and the theoretical predictions (Teich & Qian, 2003). The spatial frequency bandwidth of the same five channels is fixed to the default $h_f = 1$ octave in one set of simulations (see Figure 13, circles), or varies in proportion to the orientation bandwidth in another set (triangles). A final set of simulations varies only the spatial frequency tuning, keeping the orientation tuning fixed (squares). All other parameters are the same as in the reweighting version of the model (see Table 2). Each simulation generates the internal representations of 12,000 images counterbalanced across the various stimulus types, calculates the d's for each context and difficulty level (Equation 33), and averages them into an aggregate performance measure. As we are interested in relative improvement, we divide all results by the d' produced with default parameters.

Figure 13 shows that, indeed, the narrower the tuning bandwidths, the better the behavioral performance. However, this effect is an order of magnitude weaker than the observed learning effect. The most important aspect of Figure 13 is the extent of its vertical axis. Sharpening the tuning curves by 50% can only account for less than 10% behavioral improvement. For comparison, the average d' in our data increases more than twofold (108%), from 0.89 during the first block to 1.83 by the end of the experiment (cf. Figure 4). The retuning hypothesis is thus clearly insufficient to account for the most basic feature of our data—the size of the overall learning effect. The intuition that sharpening the representations leads to more accurate performance is qualitatively correct, and this mechanism may indeed operate in the brain and contribute to some small extent. These simulations suggest, however, that its quantitative potential is severely limited. The bulk of perceptual learning must be driven by other mechanisms.

The limitations of tuning curve sharpening ultimately reside in the spectral properties of the stimuli (cf. Figure 3). The Gabor targets have a finite spatial extent and therefore occupy a whole region[8] in Fourier space. Even with needle-sharp tuning, these stimuli still elicit a neural response spread over many units. Moreover, the targets are embedded in high-contrast noise. The recurring switch costs indicate that learning to ignore the noise is responsible for a significant fraction of the behavioral improvement. It seems that the only way to attenuate the noise via a pure retuning mechanism is to degrade the orientation selectivity of the noisy channels, which would produce obvious deficits in natural environments. External noise exclusion is fundamentally a problem of selection rather than discrimination and requires a selective reweighting mechanism rather than a representation enhancement mechanism (Dosher & Lu, 1998, 2005; Lu & Dosher, 2004).

Even if we assume, for argument's sake, that some form of representation modification could produce learning effects of the observed magnitude, it would still have to account for the recurring switch costs and the other patterns in our data. And even if it did, the modified early sensory representations would undoubtedly require reweighting of the read-out connections to subsequent areas. In conclusion, it seems that selective reweighting is not only sufficient but also necessary for perceptual learning.

The Multichannel Reweighting Model and Other Models

The new multichannel reweighting model proposed and tested in the present article embodies a number of established principles of human perception and cognition: orientation- and frequency-tuned representations, contrast gain control, weighted decision units, incremental associative learning, and intrinsic variability. The representation subsystem implements a simple "back-pocket" model of early spatial vision. The task-specific reweighting subsystem is implemented as a simple one-layer neural network that learns via a biologically plausible incremental Hebbian rule.

The representation subsystem maps the external stimulus environment onto an internal decision space. Statistical analysis of this space reveals the (nearly) linear separability of the two categories and explains why a simple one-layer classifier is sufficient to account for the perceptual learning in our task. It is of course quite possible that other tasks might require more complex (e.g., quadratic) decision bounds that the present model can handle only approximately. Whether human perceptual learning can accommodate nonlinear decision bounds is at present an important open empirical question largely unaddressed in the current literature. Two main outcomes are possible. Human observers may be able to learn complex perceptual decision spaces, in which case more complex multilayer models will be necessary. Alternatively, the behavioral evidence might indicate that the perceptual learning system is restricted to simplified solutions even for problems that require nonlinear decision bounds for optimal performance. The creation of and experimentation with critical test environments with more complex statistical structure seems an important topic for future perceptual learning research, integrating it with the nonperceptual categorization literature (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Kruschke, 1992).

The newly developed multichannel reweighting model is, by design, compatible with the generic reweighting model sketched by Dosher and Lu (1998, 1999). The new model also bears a more distant relationship to a number of other prior models and approaches (but, in our view, goes well beyond) them in its functionality and rigorous, quantitative testing (see Tsodyks & Gilbert, 2004, for an excellent review). For example, the reweighting principle is generically compatible with the *reverse hierarchy theory* (Ahissar & Hochstein, 1997, 2004), which shares the view that the early sensory representations are not to be altered unless absolutely necessary. Our model has a simpler architecture and is implemented in a fully functional system that is tested in detail. The reverse hierarchy principle is a hypothesis about patterns of learning and transfer in perceptual tasks and has been tested primarily in visual search tasks.

The present multichannel reweighting model also shares a general sensibility with the *limited-capacity* proposal of Liu and Weinshall (2000), although our model has no globally imposed capacity limitations. On the basis of experiments on motion direction discrimination, these authors conceptualize perceptual learning as a problem of finding which "measurements" of the stimulus to include in a limited-capacity sample. The measurements are divided a priori into a few discrete sets. During the initial stages of learning, the observer settles on a particular set of measurements

---

[8] The full width at half height of this region is ≈27° for our particular targets, assuming linear transduction.

that are informative for the task at hand. In a new task, the model samples preferentially from the same set, which can result in considerable time savings. This proposal is similar to the present model in that the system learns to rely on measurements with good information (signal-to-noise ratio) and filter out those with poor information; it differs in that learning is conceptualized as a limited-capacity sampling process rather than a reweighting process.

All perceptual learning models include assumptions about the sensory representations and about the learning process. Among the models most similar in spirit to our multichannel reweighting model are the models of perceptual learning for motion perception (Vaina, Sundareswaran, & Harris, 1995) and for bisection (Zhaoping, Herzog, & Dayan, 2003) and vernier hyperacuity tasks (Weiss, Edelman, & Fahle, 1993). Of course, all these models differ in the specific representations that are relevant for each task, but they all agree that perceptual learning occurs via selective reweighting. The present model bears a strong analogy to these precursors, especially that of Vaina et al. (1995), who proposed a Hebbian learning system and standard MT representations. It goes beyond most prior proposals in its focus on standard representations, its biologically plausible learning rule, its fully functional implementation that works on the actual stimulus images, and in rigorous quantitative comparison with empirical data.

All simulations reported in this article used a pure Hebbian learning rule (Equation 12). We also experimented with a member of the class of error-correcting rules—*contrastive Hebbian* (Hinton, 1989; Movellan, 1990; O'Reilly & Munakata, 2000). In our one-layer network, it is equivalent to the *delta rule* of the original perceptron (Rosenblatt, 1958). The error-correcting version of the multichannel reweighting model is also statistically driven and thus reproduces the switch costs and other effects stemming from the statistics of the stimulus distribution (Petrov, Dosher, & Lu, 2003). However, it cannot account for the slight but persistent asymmetry in favor of the context-congruent response in the human data. The error-correcting rule with perfectly counterbalanced feedback quickly eliminates any response asymmetry in the model responses unless certain restricting assumptions are made.

The Role of Feedback: A Prediction

Can the multichannel reweighting model learn in the absence of external feedback? The present experimental data are collected with feedback, and Equation 12 uses the correct response to determine the sign of each weight update. Nonetheless, *self-generated* feedback can be almost as effective as external feedback in this task (and many other tasks as well; Petrov & Anderson, 2005). As the accuracy of the model (and that of human observers) is significantly above chance at all times, including the very early trials, the model's own response is a reliable proxy for the correct response. Additional simulations indicate that such *self-supervised* Hebbian learning drives the weights to the same optimal solution in the long run. This prediction was verified in a subsequent behavioral experiment (Petrov, Dosher, & Lu, 2004, 2005).

The multichannel reweighting model is thus able to learn without external feedback under appropriate performance boundary conditions. Whenever the initial weight vector is in the right basin of attraction, reinforcing the model's own policy converges on the optimal solution. This analysis explains the pattern of earlier reports of perceptual learning without feedback (e.g., Fahle &

Edelman, 1993; Herzog & Fahle, 1997). Aggregate feedback at the end of each block of trials can help keep the participants motivated, thereby raising the overall proportion correct and the speed of learning. One possible mechanism is that feedback affects the release of catecholamines by the brain stem, which in turn modulates the gain of the neuronal activation function in the neocortex (Cohen, Aston-Jones, & Gilzenrat, 2004; Usher, Cohen, Servan-Schreiber, Rajkowski, & Aston-Jones, 1999). The higher gain augments the activity of the units that are already activated and further suppresses the activity of the units that are already being inhibited. This improves the behavioral performance, provided that it is already above chance (Servan-Schreiber, Printz, & Cohen, 1990). Conversely, when the initial weights are not in the right basin of attraction, a self-reinforcing failure to learn can result (McClelland, 2001). An interesting prediction is that the stimulus presentation schedule must include a "critical number" of easily discriminable stimuli during the early trials to ensure that the self-reinforcing learning algorithm converges to the correct partitioning of the stimulus space.

Neural Substrate

What is the neural substrate of the task-specific subsystem? We do not at present have a clear answer to this question. A large body of evidence in the domain of category learning suggests at least two separate categorization systems in the human brain: an *explicit system* for tasks with verbalizable decision rules and an *implicit system* for tasks whose rules are difficult to verbalize (see Ashby et al., 1998; Ashby & Ell, 2001, for reviews). The instructions in perceptual learning studies typically explicate a verbal response "rule." However, the stimuli are often presented briefly or in low contrast, and the tasks often require pooling of information across multiple stimulus dimensions that may be difficult to label and access consciously. Perceptual learning tends to be slow and statistically driven. There is also evidence that global amnesic patients can improve performance in a vernier discrimination task and retain the improvement when retested after 1 week (Fahle & Daum, 2002). Perceptual learning thus seems much more related to the implicit categorization system than to the explicit system, without overlapping fully with either one of them.

Several researchers have advanced the hypothesis that the basal ganglia and, in particular, the caudate nucleus in the dorsal striatum mediate a form of learning in which stimulus–response associations or habits are incrementally acquired (see Ashby et al., 1998; Packard & Knowlton, 2002; Squire, 1992; Wickens, 1997, for reviews). The striatum receives inputs from virtually all regions of the cerebral cortex, and many corticostriatal pathways are excitatory and topographically organized (McGeorge & Faull, 1989). This anatomy is entirely consistent with the hypothesis that one function of the basal ganglia is to form arbitrary S-R associations and select among competing alternatives. Whether this system is relevant for perceptual learning remains to be studied.

A number of cortical and subcortical regions must work together during the perceptual learning process. The early visual cortices (V1, V2) represent the stimulus. The specificity of learning suggests that these early cortices are involved in some kind of plasticity—either directly or, as we argue here, in the connections originating from them. If the selective reweighting hypothesis is correct, practicing a task adjusts the connections to other cortical and/or subcortical areas. This reweighting might occur in the

connections to the very next processing stages in visual analysis (V4, MT) or further along (IT, . . .). The prefrontal cortex may be involved both in maintaining task set and in monitoring feedback and bias over trials, although we do not believe that it is the locus of the new task associations (Cohen et al., 2004; Miller & Cohen, 2001). Finally, it seems likely that a reward center mediating the effects of external or self-generated feedback is involved as well (e.g., Cohen, Braver, & Brown, 2002; Montague, Hyman, & Cohen, 2004; Usher et al., 1999). These speculations will require much further study to provide a fully embodied model of perceptual learning.

Conclusions

In this article we present a critical test case of perceptual learning in two alternating training contexts that demonstrably utilize both the same (or highly overlapping) stimulus representations and the same (or highly overlapping) read-out connections. On the basis of an earlier observer analysis of a simple perceptual task within the framework of the perceptual template model, it was argued that perceptual learning in many cases is accomplished by selective reweighting of the inputs from stable, unchanging, stimulus representations (Dosher & Lu, 1998, 1999). A review of the literature suggests that only a few reported studies engage overlapping representations and connection structures between training and test that pose a legitimate challenge to reweighting. The present experiment is explicitly designed with these specifications in mind. It provides a complex pattern of behavioral regularities, including (a) consistent general improvements over a lengthy training period, (b) strong grounding effects of the target contrast, (c) recurring switch costs, and (d) a complex interaction between context congruency and target contrast. We present a fully functioning implementation of a multichannel reweighting model (multichannel perceptual template model), consisting of a standard implementation of early frequency- and orientation-specific representations coupled with incremental, Hebbian reweighting of the connections to a decision unit. The qualitative and quantitative adequacy of the implemented multichannel reweighting model provides an existence proof of the adequacy of the selective reweighting hypothesis. Furthermore, extensive analyses suggest that selective reweighting is not only sufficient but also necessary for perceptual learning.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A, 2,* 284–299.

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

Ahissar, M., & Hochstein, S. (1996). Learning pop-out detection: Specificities to stimulus characteristics. *Vision Research, 36,* 3487–3500.

Ahissar, M., & Hochstein, S. (1997, May 22). Task difficulty and the specificity of perceptual learning. *Nature, 387,* 401–406.

Ahissar, M., & Hochstein, S. (1998). Perceptual learning. In V. Walsh & J. Kulikowski (Eds.), *Perceptual constancy: Why things look as they do* (pp. 455–498). Cambridge, MA: Cambridge University Press.

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8,* 457–464.

Ahissar, M., Laiwand, R., Kozminsky, G., & Hochstein, S. (1998). Learn-

ing pop-out detection: Building representations for conflicting target–distractor relationships. *Vision Research, 38,* 3095–3107.

Ahumada, A. J., & Watson, A. B. (1985). Equivalent-noise model for contrast detection and discrimination. *Journal of the Optical Society of America A, 2,* 1133–1139.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Ashby, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Erlbaum.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442–481.

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5,* 204–210.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

Ball, K., & Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vision Research, 27,* 953–965.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience, 2,* 32–48.

Blakemore, C., & Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental Brain Research, 15,* 439–440.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Buonomano, D. V., & Merzenich, M. M. (1998). Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience, 21,* 149–186.

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981, October 2). Efficiency of human visual signal discrimination. *Science, 214,* 93–94.

Cannon, M. W., & Fullenkamp, S. C. (1991). Spatial interactions in apparent contrast: Inhibitory effects among grating patterns of different spatial frequencies, spatial positions, and orientations. *Vision Research, 31,* 1985–1998.

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

Chino, Y. M., Kaas, J. H., Smith, E., Langston, A. L., & Cheng, H. (1992). Rapid reorganization of cortical maps in adult cats following restricted deafferentation in the retina. *Vision Research, 32,* 789–796.

Chubb, C., Sperling, G., & Solomon, J. A. (1989). Texture interactions determine perceived contrast. *Proceedings of the National Academy of Sciences, USA, 8,* R33–R76.

Cohen, J. D., Aston-Jones, G., & Gilzenrat, M. S. (2004). A systems-level perspective on attention and cognitive control: Guided activation, adaptive gating, conflict monitoring, and exploitation versus exploration. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (pp. 71–90). New York: Guilford Press.

Cohen, J. D., Braver, T. S., & Brown, J. W. (2002). Computational perspectives on dopamine function in prefrontal cortex. *Current Opinion in Neurobiology, 12,* 223–229.

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

Crist, R. B., Li, W., & Gilbert, C. D. (2001). Learning to see: Experience and attention in primary visual cortex. *Nature Neuroscience, 4,* 519–525.

Darian-Smith, C., & Gilbert, C. D. (1994, April 21). Axonal sprouting accompanies functional reorganization in adult cat striate cortex. *Nature, 368,* 737–740.

Das, A. (1997). Plasticity in adult sensory cortex: A review. *Network: Computation in Neural Systems, 8,* R33–R76.

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

De Valois, R. L., Yund, E. W., & Hepler, H. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research, 22,* 531–544.

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

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

Dosher, B. A., & Lu, Z.-L. (2005). Perceptual learning in clear displays optimizes perceptual expertise: Learning the limiting process. *Proceedings of the National Academy of Sciences, USA, 102,* 5286–5290.

Eysel, U. T. (2002). Plasticity of receptive fields in early stages of the adult visual system. In M. Fahle & T. Poggio (Eds.), *Perceptual learning* (pp. 43–65). Cambridge, MA: MIT Press.

Fahle, M. (1997). Specificity of learning curvature, orientation, and vernier discriminations. *Vision Research, 37,* 1885–1895.

Fahle, M., & Daum, I. (2002). Perceptual learning in amnesia. *Neuropsychologia, 40,* 1167–1172.

Fahle, M., & Edelman, S. (1993). Long-term learning in vernier acuity: Effects of stimulus orientation, range, and feedback. *Vision Research, 33,* 397–412.

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

Ferster, D., & Miller, K. D. (2000). Neural mechanisms of orientation selectivity in the visual cortex. *Annual Review of Neuroscience, 23,* 441–471.

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

Fiorentini, A., & Berardi, N. (1980, September 4). Perceptual learning specific for orientation and spatial frequency. *Nature, 287,* 43–44.

Ghose, G. M., Yang, T., & Maunsell, J. H. R. (2002). Physiological correlates of perceptual learning in monkey V1 and V2. *Journal of Neurophysiology, 87,* 1867–1888.

Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron, 31,* 681–697.

Gilbert, C. D., & Wiesel, T. N. (1992, March 12). Receptive field dynamics in adult primary visual cortex. *Nature, 356,* 150–152.

Gonzalez, R. C., & Woods, R. E. (1992). *Digital image processing.* Reading, MA: Addison Wesley.

Graham, N., & Sutter, A. (2000). Normalization: Contrast-gain control in simple (Fourier) and complex (non-Fourier) pathways of pattern vision. *Vision Research, 40,* 2737–2761.

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks, 1,* 17–61.

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

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

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation, 1,* 143–150.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology, 160,* 106–154.

Jenkins, W. M., Merzenich, M. M., Ochs, M. T., Allard, T., & Guic-Robles, E. (1990). Functional reorganization of primary somatosensory cortex in adult owl monkeys after behaviorally controlled tactile stimulation. *Journal of Neurophysiology, 62,* 82–104.

Jones, E. G. (2000). Cortical and subcortical contributions to activity-dependent plasticity in primate somatosensory cortex. *Annual Review of Neuroscience, 23,* 1–37.

Kaas, J. H., Krubitzer, L. A., Chino, Y. M., Langston, A. L., Polley, E. H.,

& Blair, N. (1990, April 13). Reorganization of retinotopic cortical maps in adult mammals after lesions of the retina. *Science, 248,* 229–231.

Karni, A. (1996). The acquisition of perceptual and motor skills: A memory system in the adult human cortex. *Cognitive Brain Research, 5,* 39–48.

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, USA, 88,* 4966–4970.

Knutsson, H., & Granlund, G. H. (1983). Texture analysis using two-dimensional quadrature filters. In *1983 IEEE computer society workshop on computer architecture for pattern analysis and image database management* (pp. 206–213). Silver Spring, MD: IEEE Computer Society.

Koyama, S., Harner, A., & Watanabe, T. (2004). Task-dependent changes of the psychophysical motion-tuning functions in the course of perceptual learning, *Perception, 33,* 1139–1147.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Li, W., Piëch, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience, 7,* 651–657.

Liu, Z., & Weinshall, D. (2000). Mechanisms of generalization in perceptual learning. *Vision Research, 40,* 97–109.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Lu, Z.-L., Chu, W., Lee, S., & Dosher, B. A. (2005). Independent perceptual learning in monocular and binocular motion systems. *Proceedings of the National Academy of Sciences, USA, 102,* 5624–5629.

Lu, Z.-L., & Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research, 38,* 1183–1198.

Lu, Z.-L., & Dosher, B. A. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *Journal of the Optical Society of America A* [Special Issue], *16,* 764–778.

Lu, Z.-L., & Dosher, B. A. (2000). Spatial attention: Different mechanisms for central and peripheral temporal precues? *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1534–1548.

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

Lu, Z.-L., & Sperling, G. (1999). Second-order reversed phi. *Perception & Psychophysics, 61,* 1075–1088.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Marčelja, S. (1980). Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America, 70,* 1297–1300.

The MathWorks. (1999). *MATLAB user's guide.* Natick, MA: Author.

McClelland, J. L. (2001). Failures to learn and their remediation: A Hebbian account. In J. L. McClelland & R. S. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives* (pp. 97–121). Mahwah, NJ: Erlbaum.

McGeorge, A. J., & Faull, R. L. M. (1989). The organization of the projection from the cerebral cortex to the striatum in the rat. *Neuroscience, 29,* 503–537.

Medin, D. L., & Shaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202.

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

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004, October 14). Computational roles for dopamine in behavioural control. *Nature, 431,* 760–767.

Movellan, J. R. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1989 connectionist models summer school* (pp. 10–17). San Mateo, CA: Morgan Kaufman.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial and temporal contrast sensitivity of neurones in areas 17 and 18 of the cat's visual cortex. *Journal of Neurophysiology, 283,* 101–120.

Nelson, S., Toth, L., Sheth, B., & Sur, M. (1994, August 5). Orientation selectivity of cortical neurons during intracellular blockade of inhibition. *Science, 265,* 774–777.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Ohzawa, I., Sclar, G., & Freeman, R. D. (1982, July 15). Contrast gain control in the cat visual cortex. *Nature, 298,* 266–268.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience.* Cambridge, MA: MIT Press.

Packard, M. G., & Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience, 25,* 563–593.

Pelli, D. G., & Zhang, L. (1991). Accurate control of contrast on microcomputer displays. *Vision Research, 31,* 1337–1350.

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,* 383–416.

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2003). A computational model of perceptual learning through incremental channel reweighting predicts switch costs in non-stationary contexts [Abstract]. *Journal of Vision, 3*(9), 670a.

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2004). Comparable perceptual learning with and without feedback in non-stationary context: Data and model [Abstract]. *Journal of Vision, 4*(8), 306a.

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2005). *Perceptual learning without feedback in non-stationary contexts: Data and model.* Manuscript submitted for publication.

Poggio, T., Fahle, M., & Edelman, S. (1992, May 15). Fast perceptual learning in visual hyperacuity. *Science, 256,* 1018–1021.

Pollen, D. A., & Ronner, S. (1981, June 19). Phase relationship between adjacent simple cells in the visual cortex. *Science, 212,* 1409–1411.

Regan, D., & Beverley, K. I. (1985). Postadaptation orientation discrimination. *Journal of the Optical Society of America A, 2,* 147–155.

Recanzone, G. H., Merzenich, M. M., Jenkins, W. M., Grajski, K. A., & Dinse, H. R. (1992). Topographic reorganization of the hand representation in cortical area 3b of owl monkeys trained in a frequency-discrimination task. *Journal of Neurophysiology, 67,* 1031–1056.

Recanzone, G. H., Schreiner, C. E., & Merzenich, M. M. (1993). Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *Journal of Neuroscience, 13,* 87–104.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65,* 386–408.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Exploration in the microstructure of cognition: Vol. I. Foundations* (pp. 45–76). Cambridge, MA: MIT Press.

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

Schiltz, C., Bodart, J. M., Dubois, S., Dejardin, S., Michel, C., Roucoux, A., et al. (1999). Neuronal mechanisms of perceptual learning: Changes in human brain activity with training in orientation discrimination. *NeuroImage, 9,* 46–62.

Schmid, L. M., Rosa, M. G., Calford, M. B., & Ambler, J. S. (1996). Visuotopic reorganization in the primary cortex of adult cats following monocular and binocular retinal lesions. *Cerebral Cortex, 6,* 388–405.

Schoups, A. A., & Orban, G. A. (1996). Interocular transfer in perceptual learning of a pop-out discrimination task. *Proceedings of the National Academy of Sciences, USA, 93,* 7358–7362.

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

Schoups, A. A., Vogels, R., Qian, N., & Orban, G. A. (2001, August 2). Practising orientation identification improves orientation coding in V1 neurons. *Nature, 412,* 549–553.

Schwartz, S., Maquet, P., & Frith, C. (2002). Neural correlates of perceptual learning: A functional MRI study of visual texture discrimination. *Proceedings of the National Academy of Sciences, USA, 99,* 17137–17142.

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990, August 24). A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science, 249,* 892–895.

Shiu, L-P., & Pashler, H. (1992). Improvement in line orientation discrimination is retinally local but dependent on cognitive set. *Perception & Psychophysics, 52,* 582–588.

Somers, D. C., Nelson, S. B., & Sur, M. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. *Journal of Neuroscience, 15,* 5448–5465.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195–231.

Teich, A. F., & Qian, N. (2003). Learning and adaptation in a recurrent model of V1 orientation selectivity. *Journal of Neurophysiology, 89,* 2086–2100.

Tolhurst, D. J., & Thompson, P. G. (1975). Orientation illusions and after-effects: Inhibition between channels. *Vision Research, 15,* 967–972.

Tsodyks, M., & Gilbert, C. (2004, October 14). Neural networks and perceptual learning. *Nature, 431,* 775–781.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999, January 22). The role of locus coeruleus in the regulation of cognitive performance. *Science, 283,* 549–554.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108,* 550–592.

Vaina, L. M., Belliveau, J. W., des Roziers, E. B., & Zeffiro, T. A. (1998). Neural systems underlying learning and representation of global motion. *Proceedings of the National Academy of Sciences, USA, 95,* 12657–12662.

Vaina, L. M., Sundareswaran, V., & Harris, J. G. (1995). Learning to ignore: Psychophysics and computational modeling of fast learning of direction in noisy motion stimuli. *Cognitive Brain Research, 2,* 155–163.

Vogels, R., & Orban, G. A. (1985). The effect of practice on the oblique effect in line orientation judgments. *Vision Research, 25,* 1679–1687.

Watanabe, T., Náñez, J. E., & Sasaki, Y. (2001, October 25). Perceptual learning without perception. *Nature, 413,* 844–848.

Weinberger, N. M., Javid, R., & Lepan, B. (1993). Long-term retention of learning-induced receptive field plasticity in the auditory cortex. *Proceedings of the National Academy of Sciences, USA, 90,* 2394–2398.

Weiss, Y., Edelman, S., & Fahle, M. (1993). Models of perceptual learning in vernier hyperacuity. *Neural Computation, 5,* 695–718.

Wickens, J. (1997). Basal ganglia: Structure and computations. *Network: Computation in Neural Systems, 8,* R77–R109.

Wilson, H. R. (1986). Responses of spatial mechanisms can explain hyperacuity. *Vision Research, 17,* 941–947.

Yang, T., & Maunsell, J. H. R. (2004). The effect of perceptual learning on neuronal responses in monkey visual area V4. *Journal of Neuroscience, 24,* 1617–1626.

Zhang, K., Ginzburg, I., McNaughton, B., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *Journal of Neurophysiology, 79,* 1017–1044.

Zhaoping, L., Herzog, M. H., & Dayan, P. (2003). Quadratic ideal observation and recurrent preprocessing in perceptual learning. *Network: Computation in Neural Systems, 14,* 233–247.

Appendix A

Detailed Description of the Representation Subsystem

The representation subsystem is implemented by a MATLAB function converting the 64×64 input image $I(x, y)$ into a 7×5 matrix of non-negative activations $A(\theta, f)$, as outlined in Figure 6. All software is available from the authors and is also available online at http://www.socsci.uci.edu/~apetrov/

First, the image is convolved in parallel with a set of 140 *receptive fields* $RF_{\theta, f, \phi}$ (Equation 19). Each RF is a Gabor function with orientation $\theta \in \{0, \pm 15, \pm 30, \pm 45\}$, spatial frequency $f \in \{1, 1.4, 2, 2.8, 4\}$, and phase $\phi \in \{0, 90, 180, 270\}$. All convolutions are implemented via the fast Fourier transform, with zero-padding to avoid wrap-around artifacts. The filtered images are then rectified by the half-squaring operator $[\cdot \cdot \cdot]^2_+$ illustrated in Figure 6C. The resulting *phase-sensitive maps* $S(x, y, \theta, f, \phi)$ can be interpreted as activation patterns across a large retinotopic population of "simple cells" (see Figure 6D; Heeger, 1992).

$$S(x, y, \theta, f, \phi) = [RF_{\theta, f, \phi}(x, y) \otimes I(x, y)]^2_+. \quad (19)$$

The size and shape of the receptive fields in the image domain determine the corresponding tuning bandwidth in the Fourier domain. The tuning parameters are the same for all cells in the main simulations: full-width at half-height $h_\theta = 30°$ along the orientation dimension and $h_f = 1$ octave along the frequency dimension. (The simulations for the General Discussion involve individualized parametrization of select channels.) These are representative values for parafoveal simple cells in macaque striate cortex (De Valois et al., 1982, reported in Figure 4.11 of De Valois & De Valois, 1988) and correspond to elliptical RFs with axis ratio $\approx 1.35 : 1$ and size inversely proportional to f as illustrated in Figure 6B. Concretely, the standard deviation of the Gaussian envelope is 0.72, 0.51, 0.36, 0.25, and 0.18° of visual angle in the direction parallel to the grating and 0.53, 0.37, 0.26, 0.18, and 0.13° in the perpendicular direction, respectively.

Adding the four spatial phases together produces phase-invariant *energy maps* $E(x, y, \theta, f)$ via the identity $\sin^2 \phi + \cos^2 \phi = 1$ (Equation 20; Adelson & Bergen, 1985). Similar invariance can be obtained without squaring by pooling over a large number of randomly distributed phases (Hubel & Wiesel, 1962).

$$E(x, y, \theta, f) = \sum_\phi S(x, y, \theta, f, \phi) \quad (20)$$

$$C(x, y, \theta, f) = E(x, y, \theta, f)/(s^2 + N(f)). \quad (21)$$

The response normalization is implemented at this stage[9] (Equation 21). Following Heeger (1992), the energy maps are divided by a normalization term N representing shunting inhibition from a *normalization pool* of inhibitory interneurons. Consistent with physiological and psychophysical evidence, N is assumed to be essentially independent of orientation and modestly tuned for spatial frequency (Carandini et al., 1997; Chubb et al., 1989; Heeger, 1992). It is retinotopic, but given that the radius of our stimuli is less than 3 times the wavelength of the target Gabor patch ($\lambda = .05$ deg), simple averaging across the space provides a good approximation (Cannon & Fullenkamp, 1991). Concretely, the energy E is averaged across all orientations and spatial positions to yield a vector of 5 frequency-specific values $\mathbf{M}$ (Equation 22). These values are then mixed to implement the frequency tuning of the normalization pool $\mathbf{N}$. There is about 15% cross-talk between frequency bands separated by half octave and 5% from one octave (Equation 23).

$$M(f) = \mathbf{M} = \frac{1}{64 \cdot 64 \cdot 7} \sum_{x,y=1}^{64} \sum_{\theta=1}^{7} E(x, y, \theta, f) \quad (22)$$

$$N(f) = \mathbf{N} = \begin{pmatrix} .80 & .15 & .05 & & \\ .20 & .60 & .15 & .05 & \\ .05 & .15 & .60 & .15 & .05 \\ & .05 & .15 & .60 & .20 \\ & & .05 & .15 & .80 \end{pmatrix} \mathbf{M}. \quad (23)$$

Thus, the *phase-invariant maps* $C(x, y, \theta, f)$ in Equation 21 are normalized energy maps whose total activation is approximately constant for above-threshold stimulus contrasts. The *semisaturation constant* s^2 in Equation 21 affects the model behavior only at near-threshold contrasts. Since all stimuli in the present experiment are embedded in high-contrast noise, this parameter is set to zero in the simulations.

In the interest of parsimony, and consistent with task demands, each phase-invariant map is then pooled across space over a region comparable with the diameter of the experimental stimuli. The weighting kernel W_r in Equation 24 is a radially symmetric Gaussian scaled to sum to 1 and with full-width at half-height $h_r = 2.0°$ ($\sigma_{xy} = 0.85$). This particular choice is not critical; rerunning all simulations with $h_r = 1.0$ yields very similar fits.

$$A'(\theta, f) = \sum_{x,y} W_r(x, y) C(x, y, \theta, f) + \varepsilon_{\theta, f} \quad (24)$$

$$A(\theta, f) = F(A'(\theta, f)) \text{ where} \quad (25)$$

$$F(u) = \begin{cases} \dfrac{1 - e^{-\gamma u}}{1 + e^{-\gamma u}} A_{max} & \text{if } u \geq 0 \\ 0 & \text{otherwise} \end{cases}. \quad (26)$$

The spatial pooling reduces the number of *representational units* $A(\theta, f)$ to 35 (see Figure 6H). Although the stimuli in the experiment are presented either above or below the fixation point, this positional uncertainty is not implemented in the model. An extended version with two independent pools of representational units with receptive fields above and below the horizontal meridian would duplicate the present model. The learning rate parameter would have to be correspondingly rescaled.

Like all neuron-like elements, the representational units are noisy and have limited dynamic range. This is implemented by the *representation noise* $\varepsilon_{\theta, f}$ in Equation 24 and the sigmoidal activation function F in Equations 25 and 26. The noise is iid Gaussian with mean 0 and standard deviation $\sigma_r = 0.1$—a small value compared with the range of the external inputs (typically $0 \leq A' < 6$ in Equation 24). In other words, the variability in the representations is dominated by external rather than internal noise in the present experiment. The maximum activation parameter A_{max} is arbitrarily set to 0.5. The *gain* $\gamma = 0.8$ converts the normalized energy values to this internal activation scale. It is an important free parameter controlling the saturation in noisy contexts. (The decision unit uses the same activation function and the same γ in Equation 9.) Figure 7 illustrates the resulting representations.

---

[9] A more realistic implementation would have inhibitory interactions throughout the pathway.

Appendix B

Derivation of the Near-Optimal Weights in Equation 17

The converging connections in Figure 8 map the 35-dimensional representation space onto the 1-dimensional decision axis u. Equation 11 gives the exact probability of responding "right" given a particular representation vector $\mathbf{a}$, weight vector $\mathbf{w}$, and bias b. As the mapping is linear, it is easy to extend this result to a whole *population* of representation vectors $P_\mathbf{a}$. The mean and variance of the projection P_u are given by Equations 27 and 28, where $\langle\mathbf{a}\rangle$ denotes the mean representation and $\Sigma_\mathbf{a}$ denotes the (35×35) covariance matrix of $P_\mathbf{a}$:

$$\langle u \rangle = \mathbf{w}'\langle a \rangle - b \tag{27}$$

$$\mathrm{var}(u|P_\mathrm{a}) = \sigma^2_{u|a} = \mathbf{w}' \sum\nolimits_a \mathbf{w} + \sigma^2_d. \tag{28}$$

For homogenous stimulus populations, the distribution of the weighted average u is Gaussian to an excellent approximation (see Figure 12). The overall probability of responding "right" across all stimuli of particular target contrast, context, and congruence is thus given by Equation 29. (For the six-partite mixtures in each experimental block, the u distribution is no longer Gaussian, and a hierarchical formula must be used instead.)

$$P(\text{"Right"}|P_\mathrm{a}, \mathbf{w}, b) = \Phi\left(\frac{\mathbf{w}'\langle\mathbf{a}\rangle - b}{\sigma_{u|a}}\right). \tag{29}$$

The variance of the decision variable u comes from three sources: external noise in the stimuli, internal representation noise, and internal decision noise. As the spectral energy of the images is approximately constant and many different channels are ultimately averaged together, the decision noise ($\sigma^2_d = 0.038$) dominates the combined effect of the other two ($\mathbf{w}'\Sigma_\mathbf{a}\mathbf{w} \approx 0.005$ in the simulations). The total variance in Equation 28 is thus nearly equal for all stimulus classes.

Consider the pairwise comparison between two representation populations P_R and P_L with opposite targets of the same contrast. Due to the approximate homogeneity of variance $\sigma^2_{u|R} \approx \sigma^2_{u|L} \approx \mathbf{w}'\Sigma\mathbf{w} + \sigma^2_d$, the discriminability d' of this comparison is approximately:

$$d'(\mathbf{w}|P_R, P_L) \approx \frac{\mathbf{w}'\langle\mathbf{a}\rangle_R - \mathbf{w}'\langle\mathbf{a}\rangle_L}{\sqrt{\mathbf{w}'\Sigma\mathbf{w} + \sigma^2_d}}. \tag{30}$$

The bias b in Equation 29 is subtracted out in Equation 30 and, hence, d' depends only on the weight vector $\mathbf{w}$. The objective of the learning mechanism is to maximize d' by adjusting $\mathbf{w}$. The weight-bounding Equation 13 imposes the additional constraint that the norm of the weight vector is approximately constant: $\|\mathbf{w}\| \approx n$ (≈ 1 in the simulations). Thus, the

decision noise cannot be eliminated by indiscriminate strengthening of the bottom-up connections.

To approach this optimization problem mathematically, we project the decision noise back to the representation space using a matrix $\mathbf{X}$ such that $\mathbf{w}'\mathbf{X}\mathbf{w} = \sigma^2_d$ for all $\mathbf{w}$ on the sphere $\|\mathbf{w}\| = n$. The unique solution to this equation is a scalar multiple of the identity matrix $\mathbf{I}$. The total variance in Equation 28 can thus be rewritten as:

$$\sigma^2_u = \mathbf{w}'\Sigma\mathbf{w} + c^2\mathbf{w}'\mathbf{I}\mathbf{w} = \mathbf{w}'(\Sigma + c^2\mathbf{I})\mathbf{w}, \text{ where} \tag{31}$$

$$c = \sigma_d/\|\mathbf{w}\| = const. \tag{32}$$

In words, adding noise to the decision variable u is equivalent to adding a constant to all diagonal entries of the representation covariance matrix Σ. Any vector $\mathbf{w}^*$ that maximizes Equation 33 below also maximizes the original Equation 30. The solution to this equivalent optimization problem is well known (e.g., Ashby & Gott, 1988) and is given by Equation 34. The coefficient λ is determined from the constraint $\|\mathbf{w}\| = n$.

$$d'(\mathbf{w}|P_R, P_L) \approx \frac{\mathbf{w}'[\langle\mathbf{a}\rangle_R - \langle\mathbf{a}\rangle_L]}{\sqrt{\mathbf{w}'(\Sigma + c^2\mathbf{I})\mathbf{w}}} \tag{33}$$

$$\mathbf{w}^* = \lambda(\Sigma + c^2\mathbf{I})^{-1}[\langle\mathbf{a}\rangle_R - \langle\mathbf{a}\rangle_L] \tag{34}$$

$$\mathbf{b}^* = \mathbf{w}^*[\langle\mathbf{a}\rangle_R + \langle\mathbf{a}\rangle_L]/2 = \mathbf{w}^*\langle\mathbf{a}\rangle_{RL}. \tag{35}$$

The vector $\mathbf{m}_{RL} = \langle\mathbf{a}\rangle_R - \langle\mathbf{a}\rangle_L$ in Equation 34 connects the centroids of the two populations being discriminated. In the special case when Σ is a scalar multiple of the identity matrix (i.e., the representation distributions have spherical shape), the optimal weight vector $\mathbf{w}^*$ is collinear with $\mathbf{m}_{RL}$ and defines a minimum-distance classifier (Ashby & Gott, 1988). In the general case, the optimal weight vector rotates away from the directions of high variability. The introduction of stimulus-invariant decision noise makes the (equivalent) covariance structure more spherical and hence less rotation is needed for an optimal trade-off.

The off-diagonal elements of the matrix $(\Sigma + c^2\mathbf{I})$ are substantially smaller than the variances on the main diagonal and can be ignored to a good approximation. The inverse $(\Sigma + c^2\mathbf{I})^{-1}$ is thus approximated by a diagonal matrix of reciprocal values, which can be computed through purely local means and leads directly to Equation 17 in the main text.