Review

Reverse hierarchies and sensory learning

Merav Ahissar*, Mor Nahum, Israel Nelken and Shaul Hochstein

The Hebrew University, Jerusalem 91905, Israel

Revealing the relationships between perceptual representations in the brain and mechanisms of adult perceptual learning is of great importance, potentially leading to significantly improved training techniques both for improving skills in the general population and for ameliorating deficits in special populations. In this review, we summarize the essentials of reverse hierarchy theory for perceptual learning in the visual and auditory modalities and describe the theory’s implications for designing improved training procedures, for a variety of goals and populations.

Keywords: perceptual learning; visual system; auditory system; reverse hierarchy theory; dyslexia; anchoring deficit hypothesis

1. SCOPE OF THIS REVIEW

We begin with a description of the essential characteristics of cortical representations that are relevant for reverse hierarchy theory (RHT). We then present the essentials of RHT: dissociation between bottom-up stimulus processing and top-down perception. We further discuss the immediate implications of this concept on perception. Perceptual learning is viewed as linked with perception rather than with stimulus processing per se. This linkage has predictions regarding the specificity of learning to local stimulus parameters, the training procedure (stimulus variability) and global stimulus parameters (context). Finally, we discuss implications for two special populations, individuals with peripheral damage and individuals with difficulties in reading (dyslexics).

2. THE HIERARCHICAL NATURE OF SENSORY PROCESSING

Our sense organs dissect the incoming stimuli into their constituents that are localized along the sensory epithelium. Thus, different hair cells in the ear are best activated by vibrations in different frequency bands. Different photoreceptors in the eye are activated by electromagnetic energy at certain frequencies, falling on different positions in the retina. Similarly, different touch receptors are activated by pressure applied to different points on the skin. Consequently, each transduction element transfers information about a small, local aspect of the total energy flow.

Yet, our brain’s interpretations of the external world, expressed in our perceptual experiences, are immediately holistic and ecologically meaningful. We see and hear objects rather than their local constituents. Thus, we immediately perceive a complex sound as that of a breaking glass, or of a spoken word, and automatically identify the source that emitted it. On the other hand, we have no immediate percept of the frequencies that compose these complex stimuli. Similarly, we perceive faces and houses, but do not immediately perceive the retinal position of their parts. As summarized by the Gestalt psychologists, looking outside the window, we see a forest rather than the trees composing it.

It is now largely agreed that the gap between local sensation at the peripheral sense organs and global perception is mediated by local-to-global processing hierarchies. Although the pattern of connectivity is not strictly hierarchical (Felleman & Van Essen 1991; Van Essen et al. 1992; Van Essen 2005; Hegde & Felleman 2007), its general nature, expressed at the anatomical, physiological and recently also in imaging data, reflects gradually more global representations in areas farther from the periphery, even at the single cell level. Thus, in the visual modality, which has been most intensively studied, lower level representations (beginning at the retina) are believed to extract basic, general purpose local primitives, such as oriented bars or basic colours. At the other extreme, high-level representations are closely related to our global percepts (Tong et al. 1998, 2006; Sterzer & Rees 2008). Thus, at the higher levels of the visual hierarchy, there are face-specific cells (e.g. Perrett et al. 1982; Grill-Spector et al. 2004; Kanwisher & Yovel 2006) and probably even columns (Afrat et al. 2006). The transition from low levels to high levels, i.e. from local to global, is gradual and probably contains several stages, whose nature is less understood (although see Gallant et al. 1996; Riesenhuber & Poggio 1999). Figure 1a schematically illustrates the presumed visual hierarchy as described above.

The nature of the auditory hierarchy is far less understood. However, it is commonly accepted that lower level representations selectively encode fine spectro-temporal acoustic features. Thus, at the brainstem level (in the superior olivary complex), inputs from...
the two ears are compared within narrow frequency bands and with microsecond resolution (Blauert 1987; Yin & Chan 1990; Batra et al. 1997a,b; Jiang et al. 1997; Palmer et al. 2000). On the other hand, higher levels integrate across time and frequency and form more abstract, spectro-temporally broader, categories (Griffiths et al. 2000; Zatorre & Belin 2001; Warren & Griffiths 2003; Nelken 2004; Zatorre et al. 2004; Las et al. 2005; Wang et al. 2005; Winer et al. 2005; Chechik et al. 2006; Nelken & Ahissar 2006). One of these higher level representations is believed to be the phonological representation, which underlies human speech perception (e.g. Scott & Johnsrude 2003; Liebenthal et al. 2005; Scott 2005). Figure 1b schematically illustrates this concept. The nature of the hierarchies in other modalities (e.g. touch) is beyond the scope of this review, which focuses on the visual and the auditory modalities.

Two important and well-documented characteristics of this hierarchy are relevant for our subsequent discussion. First, it has a convergence–divergence pattern. Thus, on the one hand, physically different stimuli, whose low-level representations are very different, may belong to either the same external object or similar ones, which belong to the same high-level category. This convergence creates an essential generalization (e.g. a single object is perceived as same whether near or far), but at the cost of physical resolution. On the other hand, the pattern of divergence is such that physically similar stimuli, whose low-level representations are similar, may belong to different high-level categories (e.g. two acoustically similar sounds may instantiate the separate phonological categories of /ba/ and /wa/; see Nelken & Ahissar 2006). Second, there are massive feedback connections (Maunsell & Van Essen 1983; Bajo & Moore 2005; Bajo et al. 2006) that do not strongly affect receptive field properties at the level of single cells, and whose functional importance is not well understood. Both characteristics (convergence–divergence and feedback connections) are widely accepted as basic assumptions in models of hierarchical processing (e.g. the interactive activation model of McClelland & Rumelhart 1981; Rumelhart & McClelland 1982) and are also relevant for RHT, as detailed below.

3. REVERSE HIERARCHY THEORY AND IMMEDIATE PERCEPTION

(a) On the limits of (immediate) conscious perception

The reverse hierarchy theory is a concept that aims to link between the hierarchies of processing and the dynamics of perception. It was initially developed for the visual modality (Ahissar & Hochstein 1997, 2004; Hochstein & Ahissar 2002), and was later extended to the auditory modality (Nelken & Ahissar 2006; Nahum et al. 2008). It proposes that, by default, rapid perception is based on high-level representations alone. This simple assumption yields several counterintuitive implications.

First, it implies that our typical perceptual experiences (i.e. our conscious perception) reflect only the information stored at higher levels. Thus, if high levels are global, abstract and represent the ‘gist’ of ecologically relevant objects and events, it is only this gist that will be immediately experienced. Therefore, when we see a house, we immediately tag it as a house, but are not able, without further scrutiny (i.e. without the reverse hierarchy return to low-level representations), to accurately experience (and consciously note) its fine details (although accessing crude components is easier, since they are also represented at high levels). From a physiological perspective, this constraint stems from the convergence–divergence pattern described above: activation of a high-level population denoting a specific perceptual category may be the result of many, not necessarily similar, low-level activation patterns. Thus, when a specific population denoting an object category (e.g. car) is activated, we can immediately tag it as a car even though we have no immediate access to the details of its spatial components. From an ecological perspective, this limitation
is a by-product of the need to generalize across different instances of the same object or even of similar objects.

A simple example is our limited ability to see and draw the spatial characteristics of an object. When drawing two cups, for example, one close by and one at double the distance from us, we draw both at roughly the same size, since that is what we perceive. Yet, the retinal image of the closer cup is much larger than that of the distant one. According to the RHT account, our perception reflects higher level representations, which are the same (or similar) for the two retinally different cups. If we want other individuals who observe our drawing to have an immediate depth perception, it has to have the retinal (perspective) characteristics that the cups induce. However, the lower level representations that retain these properties are not easily accessible.

In the auditory modality, the consequences of our immediate percepts being exclusively based on high-level representations are even more dramatic. Here too, different low-level activations can activate the same high-level representation (Nahum et al. 2008). Thus, when we listen to music, we can identify the tune and tag it. However, typically we cannot explicitly access the information that is implicitly used for this identification (e.g. to decide whether two subsequent notes are going ‘up’ or ‘down’). When we hear a speech sound (e.g. syllable), we are not consciously aware of the formant frequencies and transitions that composite it, but have no problem categorizing it or even repeating it.

(b) On the limits of implicit perception

Does immediate categorization benefit from all low-level information, even though we are not aware of these details? The RHT’s approach to this question is ecological. Local details are retained separately along the bottom-up hierarchy when they discriminate between basic high-level categories (e.g. eyes versus moles in faces). On the other hand, even crude information is merged (e.g. across retinal positions) when combined into the same category along the processing hierarchy. According to RHT, the limitations of immediate vision are revealed when within-category discriminations, particularly non-practised ones, are required. This ecologically driven concept may concur with the view that low spatial frequencies are processed faster than high spatial frequencies (Bar 2003; Bar et al. 2006) and help form some top-down category-level expectation. However, RHT asserts that this will be the case only when low-level frequencies are those crucially relevant for categorical distinctions. Thus, if we practise discriminations between objects that differ only at high frequencies, the bottom-up hierarchy will retain their separation at high, immediately accessible levels. RHT may also concur with Bayesian views of perception, as gradually refining inferences regarding the external stimuli, from coarse to fine (Hegde 2008). However, it specifically asserts that ‘coarse’ refers to representations at high levels, which were formed to represent ecologically important distinctions.

To assess this concept in the auditory modality, we (Nahum et al. 2008) conducted an experiment measuring speech perception in noise. We asked whether listeners who need to discriminate between words that are phonologically similar (and hence do not belong to clearly distinct perceptual categories, e.g. /amid/ vs /tamid/), while making semantic decisions can benefit from all low-level information (figure 2a). Although the words we used sounded similar (i.e. were not categorically different), they were very different acoustically, and therefore well separated at low levels of the auditory hierarchy (see discussion and supporting information in Nahum et al. 2008). We reasoned that the requirement for comprehension mimics daily experiences by forcing listeners to base their perception on high-level representations. As shown in figure 2b, under these conditions, discrimination ability was substantially poorer than the performance they could have achieved had they used all low-level information. Thus, there is a trade-off between perceptual generalization to more abstract conceptual categories and accuracy in discrimination within categories.

The complementary question is whether perception can fully use low-level information when discriminating between phonologically non-overlapping words (e.g. /tamid/ vs /shalom/) that denote different perceptual categories. According to RHT, in that case, high levels retain low-level segregations (here in addition the low-level information is similar to that present for the phonologically similar pair, compare empty bars in figure 2b,c). Indeed, as shown in figure 2c, under these conditions, performance was optimal (i.e. used all low-level information).

This pattern of results shows that there is a perceptual cost for the convergence–divergence structure of the ascending sensory pathways in conjunction with the exclusive accessibility to high levels. This cost is expressed in the case of discriminating between phonologically similar words, where high-level representations abstract and generalize over the low-level representations. However, the case of phonologically different words seems more common ecologically, since in the context of a conversation, words are semantically and syntactically related, and the listener has a good prediction of the words between which he or she has to discriminate. The listener would therefore need to identify which of several semantically related words was heard, but typically these words do not sound alike. For example, we are more likely to need to make discriminations between /night/ and /day/, which are semantically but not perceptually close, rather than between /bay/ and /day/ which are perceptually, but not semantically close. Hence, under these conditions, loaded with top-down semantic and syntactic expectations, crude categorical discriminations probably suffice, leading to optimal discriminations, even when these are based only on high-level representations.

4. PERCEPTION WITH SCRUTINY

According to RHT, immediate perception is limited when low-level resolution is not retained at high representation levels. Yet, under special conditions, low-level information may be accessed directly. If a higher level population is roughly informative, but not sufficient for successful performance, a backward
search may be initiated to locate its most informative inputs. Such a backward search requires time and/or repetitions. A single long-duration stimulus that clearly presents the cues that need be resolved (e.g. ‘eureka’ in Ahissar & Hochstein (1997); or a clear speech signal that clarifies noisy or impoverished speech in sine wave speech in Remez et al. (1981, 2001) and Sheffert et al. (2002); or noise-vocoded speech in, e.g. Hervais-Adelman et al. (2008)) may suffice for a successful backward search. However, under typical stimulation conditions, in which the signal-to-noise ratio is not very good, a successful backward search requires repetition of the same stimuli in a sequence. When repeated stimuli are used, the backward search can be successful in identifying a more informative input population, resulting in better behavioural performance. However, the expected drawback in accessing low-level information is the loss of concurrent access to high-level information. As a consequence, we temporarily lose the benefit of global and ecologically meaningful perception, afforded by higher level representations. Thus, for example, we cannot have accurate within-category discrimination, with concomitant semantic processing.

To test these predictions, we again employed experiments requiring speech perception in noise, as described above (figure 2). However, we now omitted the comprehension requirement, by asking subjects only to repeat the word presented in noise (see Nahum et al. 2008 for details). We reasoned that now listeners...
would be able to implement a successful backward search and effectively use low-level information to improve discrimination. As shown in figure 3, this was indeed the case. When mere repetition was required, the low-level information used for the discrimination matched the maximal acoustic information of low-level representations, for both phonologically similar and different words. According to RHT, in the phonologically similar pair, the mere repetition condition, as opposed to the semantic condition, allowed subjects to address low-level representations by following reverse hierarchy pathways over the many repetitions of the task.

5. WHAT DOES TRADITIONAL PSYCHOPHYSICS MEASURE?

Although we usually encounter very complex stimuli in our everyday experiences, much of psychophysics is dedicated to assessing performance with simple stimuli. The implicit underlying assumption is that using simple stimuli enables the study of early, low-level processing, and that the understanding of this level could serve as a basis for deciphering the perception of complex objects. According to RHT, a major drawback of this approach is the incongruence between the experimenters’ attempt to assess the subjects’ access to low-level representations, by using simple stimuli, and our default use of high-level representations at all times, whether asked about a face, a word, an oriented bar, a Gabor patch or a pure tone. Specifically, RHT posits that the complexity of the stimulus does not change the default level used (Hochstein & Ahissar 2002).

However, another characteristic of many psychophysical studies is the measurement of performance using a set of repeated stimuli around a fixed reference stimulus. For example, in auditory frequency discrimination tasks, performance is usually assessed around a reference, with every trial typically containing either a single stimulus around (a little below or above) a fixed frequency or two stimuli, one of which (the reference) has this fixed frequency (e.g. Harris 1948; Watson et al. 1975, 1976; Demany 1985; Botte 1995; Irvine et al. 2000). Similarly, in typical methods for assessing other basic auditory (e.g. azimuth, intensity, duration, frequency modulation) or visual (contrast, orientation, spatial frequency) discrimination abilities, stimuli are clustered around a fixed mean (Demany 1985; Demany & Semal 2002; Adini et al. 2002; Yu et al. 2002, 2004; Karmarkar & Buonomano 2003; Fitzgerald & Wright 2005; Amitay et al. 2006). This repetition typically stems from the attempt to acquire improved statistics when evaluating performance, with the assumption that resolution around different parameters may differ (e.g. with a logarithmic dependence following Weber’s law). According to RHT, this technique allows access to low-level representations and consequently has a major impact on the performer’s discrimination abilities.

Thus, in the first trial of the assessment, performance is typically based on the default, easily accessible high levels (although it, too, depends on the nature of the preceding assessments). However, with subsequent repetitions using a narrow range of stimuli, the performer gradually gains access, via a backward search, to lower level representations that are more suitable for performing such fine discrimination tasks. It therefore follows that the high-resolution (i.e. low) thresholds found in traditional psychophysics indeed assess performance based on low-level representations, and thus evaluate the maximal information available to the organism.

However, attaining such low thresholds critically depends on the assessment procedure. For example, if stimuli are randomly varying throughout the assessment, so that when stimuli are broadly jittered across trials with no consistent reference, the evaluated thresholds (i.e. perceptual ‘sharpness’ under these conditions) would be much poorer and would not reflect low-level resolution. This phenomenon (the impact of assessment protocol) was indeed observed in the early stages of systematic psychophysics. Harris (1948), for example, found that the best thresholds for auditory frequency discrimination are obtained when using a consistent cross-trial reference stimulus. Similar observations were found in the visual modality (Helson 1947, 1948; Morgan et al. 2000; Nachmias 2006). Traditionally, psychophysicists embraced the assessment procedures that used consistent stimuli across trials (‘blocks’), partly due to the resulting better thresholds. In RHT terminology, this protocol allows the evaluation of low-level thresholds.

The complementary aspect of this approach is that these fine thresholds, measured under ‘blocked’ protocols, do not characterize the information available for perception in natural contexts. These, according to RHT, depend on high-level resolution and are therefore limited, for both simple and complex stimuli, as described above.

6. PERCEPTUAL LEARNING

(a) The gradual specificity to local stimulus attributes

RHT’s account of perception implies that under daily situations, our use of high-level representations is typically quite efficient for the identification of ecologically important scenes or events. For such categorical evaluations, our conscious perception is quick and uses all relevant low-level information. This general concept was supported by a study that assessed and attempted to improve scene identification of rapidly presented stimuli (Fabre-Thorpe et al. 2001). Initial performance was impressive, but no improvement was found with further training.

Identification of complex novel combinations should naturally be learned. However, a less intuitive observation is the difficulty in discrimination between simple similar stimuli, which are already well represented within the perceptual pathways. According to RHT, in these cases, naive performance is not limited by low-level information but by its loss at the easily accessible higher levels. Practice-induced improvements of discrimination between similar stimuli are therefore the result of a gradually gained access to more informative, lower level populations (Ahissar & Hochstein 2004).
In line with the concept of top-down-driven learning, several recent studies have shown that high-level task clarity is essential for obtaining improvement. Thus, Garrigan & Kellman (2008) have found that high-level perceptual constancy rather than low-level sensory constancy is crucial for learning. Zhang et al. (2008) have found that top-down segregating cues (‘tagging’) allowed the perceptual system to dissociate between similar, sequentially presented inputs and hence to improve contrast discrimination. Once learning ‘kicks off’ and relevant inputs are strengthened, the obtained improvement is retained for a very long period (Karni & Sagi 1993; Polat et al. 2004).

Thus, according to RHT, naive (untrained) performance is based on high-level representations, whereas trained visual performance that requires fine spatial resolution is based on lower level representations. The general concept of increasing the weights of task-relevant (i.e. informative) inputs is also shared by other models of perceptual learning (e.g. Dosher & Lu 1998, 1999, 2000). RHT puts it in the broader context of perceptual hierarchies and explicit perception (see comparison in Yotsumoto & Watanabe 2008). The idea that naive perception reflects high levels, whereas trained performance with respect to local attributes is more closely related to lower level activity has received strong support from recent functional magnetic resonance imaging findings in humans (Schwartz et al. 2002; Furmanski et al. 2004; Sigman et al. 2005; Mukai et al. 2007). For example, Sigman et al. (2005) found that when observers were trained to search for a local T-shaped target, initial performance level was correlated with activity in a higher order area (lateral occipital cortex), whereas subsequent performance was correlated with activity at earlier, retinotopically organized, areas. Similarly, Furmanski et al. (2004) showed that after practicing for a month on the detection of low-contrast oriented patterns, V1 response for the practised orientations significantly increased. A monocular study of texture discrimination (Schwartz et al. 2002) showed that changes following a single intensive session of training on texture discrimination were restricted to the corresponding retinotopic area in the early visual cortex. The time scale required for reaching low levels is still not well understood. According to RHT, it depends on the difficulty (signal-to-noise ratio) of the informative low-level populations.

Since the increase in resolution is obtained through access to lower level representations, this additional improvement is expected to have the same specificity signature as that of the lower level populations that underlie it. Indeed, one of the markers of these effective training procedures, which use a narrow range of stimuli, allowing access to specific lower level populations, is the specificity of this subsequent improvement. Thus, in a range of visual (e.g. Fiorentini & Berardi 1980; Poggio et al. 1992; Shiu & Pashler 1992; Ahissar & Hochstein 1993; Fahle 1994; Sagi & Tanne 1994; Fahle et al. 1995; Soups et al. 1995; Adini et al. 2002; Schwartz et al. 2002) and auditory (Watson et al. 1976; Wright & Fitzgerald 2001; Demany & Semal 2002) tasks, improvement was found to be quite specific with respect to dimensions that are well segregated at lower but not higher levels of the processing hierarchies.

An alternative training approach to the studies described above is the use of globally complex (rather than simple impoverished) environments. Does massive training in complex contexts lead to low-level modifications? According to RHT, such an effect would depend on two factors: the first is that performance can be improved by perception reaching specific low-level populations; this would imply that naive performance was limited by lack of fine resolution. The second requirement is that these low level populations can be tracked (e.g. intertrial variability is limited, as described above).

In accordance with these predictions, Sowden et al. (2000) found that expert radiologists have better contrast sensitivity to points in X-rays than novices, indicating that a relevant lower level cue, potentially crucial for...
diagnostics, had improved, at least within their trained context. On the other hand, Pelli et al. (2006) found no evidence for specialized letter identification detectors in native compared with ad hoc alphabets. That is, even in highly trained individuals (practically all adults), the ability to identify letters is well predicted by the letters’ visual complexity, and does not seem to improve with age. Moreover, following no more than 3000 training trials with a totally new alphabet, the level of identification proficiency was similar to that of the highly trained alphabet. Nevertheless, an advantage of massive training is that perceptual learning of fine details may be very large for the massively trained than for the native compared with ad hoc alphabets. That is, even in the case of reading in context, letter identification rate is does not automatically yield specialized detectors. In the case of reading in context, letter identification rate is probably not the limiting factor for reading rate.

(b) **The impact of the training protocol**

As described above, gaining access to specific lower level populations requires a backward search for the most informative populations. Determining which population is informative requires either a repetition across a number of sequential trials or an atypically large signal-to-noise ratio. Indeed, a major prediction of RHT is that perceptual learning of fine details may not be attained without blocked presentation of the relevant cues. This implies that a successful back tracking search may not be achieved when stimulus variability is increased in such a way that consecutive stimuli are still within the same perceptual category, and hence activate the same high-level population, if they are still sufficiently different and thus activate different low-level populations. Similar variability patterns are expected to interfere with the dynamics of sharpening discriminations for both short- and long-term learning. Thus, whether the variability relates to a task-relevant (e.g. orientation in an orientation discrimination task) or task-irrelevant (spatial frequency in an orientation discrimination task) dimension from the experimenter’s perspective, its impact is expected to depend on the relationships between these dimensions at low-level representations (e.g. if changing spatial frequency will affect the most informative population for orientation). If no informative low-level population can be consistently tracked, RHT predicts that performance will be based on high-level resolution. It will therefore show only a limited degree of improvement on the one hand, and substantial generalization across low-level features on the other hand, as illustrated in figure 4. Indeed, perceptual learning studies in the visual modality that used mixed stimuli (i.e. a ‘roving’ protocol) found only limited learning or no learning at all (e.g. Adini et al. 2004; Yu et al. 2004; note that when the stimuli are mixed in the same manner repeatedly, it introduces a special case, Kuai et al. 2005).

Similar results were reported in the auditory modality, and specifically in the speech perception domain. For example, Mullenix & Pisoni (1990; Mullenix et al. 1989; see also Green et al. 1997) found that identification of words in noise was better when the same

---

**Figure 5.** Illustration of the crowding effect. Fixate on the dot (filled circles) on the left of (a,b) and try to identify the central letter on the right. (a) Isolated letter—identification is easy when only one letter is presented. (b) Crowding—identification is much harder when other letters surround the middle letter.

**Figure 6.** A schematic of the top-down cascade of learning across hierarchical layers. For simplicity, three levels of a hierarchical structure are featured. (a) Initial learning phase. Since learning follows reverse hierarchy order, the first connection strengthened is the presumed informative connection which feeds into the higher node (red thick line; connecting between nodes c and B). (b) Subsequent learning phase. With more training on specific conditions, learning proceeds backwards to the most informative lower level population, strengthening the connection between lower node 4 and middle node c (path denoted in red). As a result, the response in upper level node B is strengthened, but since node c also feeds into upper nodes A and C, their connections are also somewhat strengthened (dotted red lines). Therefore, following this phase, learning is partially transferred to higher level contexts that use the trained lower level features.
speaker was used consistently throughout the block compared with when different speakers were interleaved throughout the block. The complementary prediction was also reported: Logan et al. (1991; see also Lively et al. 1993) showed that a more variable training set led to greater generalization of the discrimination between the English consonants /r/ and /l/ for Japanese speaking subjects. Finally, Clopper & Pisoni (2004) who trained subjects to classify sentences according to dialect region, found that the group that received the more variable training was less accurate at the initial identification phase, but generalized better when subsequently tested on the classification of new speakers with new sentences.

It therefore follows that training with a narrow range of stimuli creates an expert on the narrow range trained, whereas variability in training interferes with the specific search yet provides broader training with less expertise. Thus, if one wants to train an expert on a broad range of stimuli, an effective way of doing so would be to present the broad range of stimuli in a blocked, rather than an interleaved manner. Blocked presentations provide useful expectations, which guide the backward search. Moreover, several studies indicate that, following training in a blocked manner, performance improvement transfers to the interleaved condition (e.g. Adini et al. 2004). Thus, the role of expectations is much more important for effective training than for effective expert performance.

(c) Training and transfer to untrained global stimulus parameters (context)

We have thus far discussed transfer of learning to new local stimulus parameters. An equally important question is the transfer to new aspects of trained stimuli. RHT proposes that when a new task requires the use of the same stimuli in a completely unrelated manner, learning will not transfer. This prediction has been amply verified (e.g. Greenspan et al. 1988; Ahissar & Hochstein 1993; Shiu & Pashler 1995; Fahle & Morgan 1996; Fahle 1997; Nygaard & Pisoni 1998). The other side of this reasoning is that learning is expected to transfer if the trained cue is still relevant for the new task: Webb et al. (2007) demonstrated transfer of learning in a spatial axis judgement task across both task and stimulus configurations, providing that the same axis of judgement remained relevant.

An even more challenging question theoretically, and probably more relevant for practical purposes, is the degree of learning generalization when the trained local cues are presented in a different global context. Will our trained perception identify the similarity to the training conditions and benefit from prior practice, or, alternatively when the ‘gist of the scene’ is significantly (categorically) modified, locally trained cues will not be available for perception any more? This question has hardly been addressed, either theoretically or experimentally. Since it is an extremely important one, we shall discuss few cases and interpret them from the perspective of RHT.

(i) Case I: visual learning—from letter identification to words

One of the few studied cases is the local-to-global transfer across context from improving letter identification to improving reading rate. This question is particularly relevant in peripheral reading.

Our ability to identify small crowded objects substantially decreases with increased eccentricity (distance from the centre of the visual field). The reason for this is the decrease in visual acuity with eccentricity and the increased interference of neighbouring elements (crowding; see figure 5 for an illustration). The mechanisms underlying crowding effects are not well understood, but it is quite clear that they rely, at least partially, on high-level object recognition mechanisms, beyond V1 (Levi 2008).

In contrast to accuracy per se, which cannot be improved with training (Westheimer & Truong 1988), peripheral crowding can be reduced with practice. Since peripheral reading rate is known to be limited by crowding (Levi et al. 2007), it is interesting to assess whether reduced crowding transfers to faster peripheral reading. In addition, this question has practical implications since many individuals suffer from a loss of central vision (e.g. in macular degeneration, see Ciulla et al. 1998) and could hugely benefit from peripheral reading, which is extremely difficult.

Motivated by this question, two separate studies trained subjects with normal vision to reduce their peripheral crowding using somewhat different tasks. In both studies, a set of three letters was presented at a given position in the periphery. However, in one study (Chung et al. 2004) subjects were asked to name all three letters, whereas in the other (Chung 2007) observers were asked only to name the central (most crowded) letter. Both training paradigms led to reduced crowding effects as measured by improved task performance. However, only training with the full three-letter identification task led to an increase in peripheral reading rate. Thus, in both studies, training improved the task-related cues, but only the three-letter identification task shares the same visual constraints as peripheral reading, apparently including retention of ‘separate lines’ for each of the letters composing the words. Training on identification of the central letter is likely to have functionally ‘shrunk’ the focus of attention. This pattern of limited transfer indicates that transfer crucially depends not just on use of similar visual patterns for training and assessment, but rather on use of the same ‘readout’ mechanisms and the new task being limited by the same bottlenecks as the trained one.

Another population who could benefit from reduced crowding is the large population of amblyopes (individuals with a ‘lazy eye’). Amblyopia results from optical impairments during development (either a great difference between the optics of the two eyes, or from strabismus; Williams & Harrad 2006), which reduces the visual accuracy of the amblyopic eye even after the optical limitations are corrected (if such a correction is not applied at young childhood; see recent review of Levi 2006). In addition, amblyopes suffer from crowding even in central vision, leading to slower reading rate with the amblyopic eye and the need for larger distances between print letters (Levi et al. 2007). Polat et al. (2004) conducted an intensive visual training protocol to ameliorate performance of the amblyopic eye. Observers practised on detection of
very dim Gabor patches at various spatial frequencies and orientations, in a blocked protocol (i.e. one spatial frequency and one orientation per session), and most of them significantly improved. More importantly, improvement transferred to the standard letter identification task of visual acuity, and their central crowding was also reduced. The transfer to new visual contexts and somewhat different tasks may stem from a common high-level mechanism. For example, if amblyopes typically learned to suppress the amblyopic information at a high-level stage (as high-level aspects of binocular rivalry), the guided practice may have gradually trained them to actively use this information. If this is the case, amblyopic learning should result in much broader transfer than that in the general population (e.g. Huang et al. 2008). Alternatively, transfer may be found only to tasks that use the trained cues.

Would such training increase reading rate using the amblyopic eye? To the best of our knowledge, this question has not been assessed experimentally. However, its answer will be quite revealing. If increased central crowding results from allocating fewer effective output channels to the amblyopic information, such local training may not suffice. As in the general population, only accuracy and single-crowded letter identification will improve (as described above). However, if their increased central crowding is a general outcome of a higher level bottleneck, then learning will be much broader than in the general population, and reading rate will increase.

(ii) Case II: auditory learning—from words to sentences
A related question in the auditory modality is whether training on isolated words would transfer across context, i.e. to sentences and fluent speech. In our daily communications, we are constantly, implicitly, deciphering connected speech in context. Such speech is acoustically very different from isolated words and contains additional supra-segmental cues, such as the prosody and intonation of the speaker, which provide a context in which words can be deciphered. Generalization from laboratory training on isolated words to everyday conversational context is therefore highly important not just for the normal population, but also for several populations with hearing deficits, such as individuals with cochlear implants. For these populations, such transfer may greatly facilitate everyday coping with the complex acoustic environment in which they operate.

However, as predicted by RHT for the visual domain, it seems that transfer of speech learning occurs only if the cues that would later be relevant in the broader context of the test are the same cues that limit performance on the trained stimuli. For example, Greenspan et al. (1988) studied learning and generalization of synthetic speech. The authors trained listeners to identify isolated words or sentences generated by a synthetic speech generator device. Indeed, they found no generalization for synthetic speech across context: thus, subjects who were given training on isolated words improved only on identification of isolated words, whereas subjects trained with sentences mainly improved on identification of the words within the sentences. As stated by the authors, in such synthetic speech, the cues for word boundaries (beginning and ending) are different when words are presented in isolation than when they are concatenated to produce a sentence. Moreover, in a sentence of concatenated synthetic words, unlike in a sentence of natural connected speech, supra-segmental cues are not present and cannot aid segmentation. Presumably, subjects trained on isolated words did not learn the top-down cues that are required for segmentation of the concatenated synthetic words in the absence of regular word boundary cues, whereas the cues that they did learn, which helped them identifying the isolated words, were useless in the broader context of sentence identification. By contrast, these specific cues were acquired in the sentence-trained group. Similar results were obtained in a later study by Nygaard & Pisoni (1998) who trained subjects on extracting relevant speech cues from either words or sentences of natural speech. The authors found that when subjects attended to sentence level cues during training, learning was not generalized to recognition of new isolated cues, but did generalize to identification of words within sentences. These results again support the idea that the dimension which was relevant during training was the one later transferred to the testing phase.

(iii) The conceptual RHT scheme for consistent and massive training: local specificity with global generalization
According to RHT, prolonged learning leads to increased cue specificity (given the conditions discussed above). However, at the same time, prolonged learning can lead to increased generalization across different global contexts.

With prolonged training on challenging, non-trivial conditions, improvement can gradually progress backwards to lower, more specific levels, as illustrated in figure 6. Thus, for the highly trained expert, learning is expected to modify low-level representations, and is therefore gradually more specific to local stimulus attributes. However, a complementary consequence is that these modifications feed-forward into additional higher level nodes, which were not initially activated, since learning took place in a globally different environment. Thus, expert performance will be both more locally specific (due to the reverse hierarchy) and more globally general (due to divergence of feed-forward connections), as shown in figure 6b.

Several recent studies have provided support for this concept. For example, Burk & Humes (2007) gave subjects either short (5 hours) or long (15 hours) training on the identification of words in noise, and then tested the generalization of learning to sentences, among other factors. They found that the long, but not the short training was generalized from isolated words to fluent speech. In RHT terms, it may be that the longer training enabled improvement in lower representation levels, which provide generalization across broader global contexts, as explained above. Similarly, Nishi & Kewley-Port (2007) found that training Japanese speakers on the entire set of English vowels were generalized across context, to other words and speakers, whereas training on a subset of the vowels did not. Here too, it seems that the relevant factor...
determining generalization is the degree to which the training set encompasses the entire low-level variation within the speech material.

7. IMPLICATIONS TO SPECIAL POPULATIONS
(a) Perceptual learning in the case of peripheral damage
We have so far stressed the major role and impact of top-down learning: task-related learning where the weights of the presumably relevant cues are enhanced to improve task performance (Ahissar & Hochstein 2004). However, the case of peripheral damage is different, and stresses the role of bottom-up induced modifications. Here, the input pattern changes, and the system needs to adapt to it.

Under normal conditions, central representations are continuously modified, adapting to the changes in the external environment. Such changes do not require a change in the readout mechanisms since our perception operates on a relative scale, and absolute increases or decreases of inputs do not yield different perceptual labels. Yet, in injury, this bottom-up tracking mechanism may yield misrepresentation of the external world. In this case, the re-distribution of the inputs reaching central representations results from impaired sampling of the external environment. Hence, if readout mechanisms are not updated, which is probably the case for regular bottom-up changes, as described above, these changes will be misinterpreted.

A strong such example was provided by Ramachandran et al. (1992a,b) who found that in a person whose arm was amputated, touching the face also felt like touching the phantom limb. Presumably, bottom-up activation modified the pattern of representations, and inputs from the face now also activate neural populations that were previously activated only by the arm (Ramachandran et al. 1992a,b). When the face is touched, both face representations and neighbouring representations are activated. From a bottom-up perspective, both are now activated by the face periphery. Yet, the readout mechanism was not ‘informed’ and still interprets their activation as originating from the limb, leading to the experience described above.

Individuals with false sensations due to peripheral injuries could significantly benefit from specific top-down guided learning, whose aim would be to ‘update’ the readout mechanism and avoid such confusion (see discussion in Ahissar & Ahissar 1994). Task-specific shaping of high-level readout mechanisms is exactly what happens in the first stages of perceptual learning (figure 6a), as described above.

(b) Can perceptual learning ameliorate dyslexia?
When individuals have a peripheral deficit (e.g. ambylopic amputees), it is easy to detect the source of their impaired perceptual performance. However, many populations with adequate peripheral mechanisms and at least large-scale normal brain anatomy nevertheless show impaired performance when their perception is assessed in controlled laboratory conditions. Such is the case with individuals with learning disabilities. Numerious studies have shown that many dyslexic individuals show poor performance in a range of visual (Lovegrove et al. 1980; Lehmkuhle et al. 1993; Cornelissen et al. 1995; Gross-Glenn et al. 1995; Eden et al. 1996; Stein & Walsh 1997; Ben-Yehudah et al. 2001; Ben-Yehudah & Ahissar 2004; Sperling et al. 2006) and auditory (Tallal 1980; McNally & Stein 1996; Ahissar et al. 2000; Hari & Renvall 2001; Amitay et al. 2002; France et al. 2002; Goswami et al. 2002; Mengler et al. 2005) tasks. Can these populations academically benefit from perceptual training?

A relevant question is the core deficit underlying their poor perceptual performance. A dominant hypothesis suggested a general impairment in rapid stimulus processing (Tallal 1980), perhaps specifically related to the magnocellular pathways (e.g. Stein & Walsh 1997; Stein 2001). A more recent hypothesis has suggested a general ‘noise-exclusion deficit’ (Sperling et al. 2005). Ahissar and colleagues noted that dyslexics’ deficit is task and context dependent (Banai & Ahissar 2006), and cannot be accounted for by a specific low-level impairment (Amitay et al. 2002). The dyslexics’ performance crucially depended on the degree of stimulus repetition during the assessment. Specifically, in standard psychophysical protocols measured around a repeated reference stimulus, controls significantly benefitted from the cross-trial stimulus-specific repetitions, in line with RHT (see ‘traditional psychophysics’ §5), whereas dyslexics were impaired in using these cross-trial consistencies (Ahissar et al. 2006). When no cross-trial repetition was used, controls were as poor as dyslexics. A similar deficit characterized their speech perception. This led Ahissar (2007) to suggest that a stimulus ‘anchoring deficit’ impedes a broad range of their skills, including verbal memory and consequently reading.

The anchoring deficit could be interpreted, in line with RHT, as the inability to use stimulus-specific repetitions for accessing low-level informative populations, which are important for gaining improved perceptual resolution. Hence, higher level representations keep dominating perception, even when access to low levels is beneficial. A failure in such a search may result from impaired attentional, backward search, mechanisms. This interpretation is in line with the attentional difficulties that many dyslexics have (Facoetti et al. 2000a,b; Bednarek et al. 2004; Buchholz & McKone 2004; Roach & Hogben 2007). Yet dyslexics’ automatic perceptual skills, assessed with auditory oddball paradigms using evoked response measurement while individuals are watching a silent film (the mismatch negativity wave, MMN Naatanen 1992), also seem impaired (Baldeweg et al. 1999; Kujala et al. 2003; Renvall & Hari 2003; Corbera et al. 2006; Huttunen-Scott et al. 2008); although see recent review by Bishop (2007), for inconsistencies in MMN findings for dyslexics). Thus, a second interpretation might be that dyslexics’ lower level populations have impaired resolution. However, when assessed with a frequency change detection task (i.e. a same–different paradigm) using the same stimuli, no deficit is found (Banai & Ahissar 2006). Namely, dyslexic individuals have the same frequency resolution as controls. Taken together, these data suggest that the impairment in dyslexia is the inability to retain stimulus-specific

Phil. Trans. R. Soc. B (2009)
information across intervening stimuli. Such a deficit would impair the ability to allocate consistently activated informative populations.

In order to assess whether perceptual learning can improve dyslexics' anchoring ability, and transfers to an untrained range (e.g. from tones to words), we applied a perceptual training protocol, in which subjects were trained on two-tone discrimination tasks, starting with frequency discrimination, followed by duration discrimination and other basic dimensions. Most participants improved in the trained tasks. More importantly, following training, their ability to perform verbal memory tasks that rely on repeated stimulus presentations was also improved (Banai & Ahissar submitted). These findings suggest that a shared bottleneck limits perceptual and verbal memory anchoring in dyslexia. The characteristics of this shared implicit memory mechanism are beyond the current scope of RHT; and should perhaps be incorporated into a broader version that specifies the mechanisms underlying the allocation of informative neuronal populations.

The dyslexia training study suggests that perceptual training paradigms may be useful even when the underlying bottleneck for performance does not seem to directly involve the adequacy of strictly perceptual representations. Perceptual training may be used as a highly adaptive tool that can also probe and modify basic cognitive mechanisms that are used by various levels of representations.

8. CONCLUDING REMARKS
Recent studies have suggested that perceptual training may be used as an effective tool for upgrading a broad range of abilities, including perceptual expertise, rehabilitation of abnormal perception and improving cognitive skills. The RHT attempts to account for all these cases within a unified concept that assumes a top-down-driven learning cascade. Its importance lies in yielding specific, sometimes counter-intuitive predictions. For example, the RHT had been successful in predicting the importance of the training protocol (i.e. the sequence of stimulus presentations) and its impact on the amount, rate and generalization of the resulting improvement. It naturally accounts for the importance of the similarity between the trained cues and those that need be used later, in the testing phase. A crucial question relates to the transfer across stimulus context, from global to local and vice versa. The RHT predicts that global-to-local transfer will only occur to the extent that local cues are both important (i.e. form a bottleneck to performance) and accessible (depending on the consistency of the cues throughout training). Local-to-global transfer would only occur following substantial training, and only for cues that are also informative in the novel global conditions. In this review, we have shown that these general predictions account for a broad range of perceptual learning studies in the visual and auditory modalities.

Although RHT has been useful in predicting and providing specific guidelines for effective training, the job is far from being complete. Deriving more detailed concepts of training is of great importance given the huge potential of perceptual learning to different populations on the one hand and the large sensitivity to the training procedures on the other hand.

We thank the Israel Science Foundation and NIH grant member 2R01 DC004855 from NIDCD for supporting this research.

REFERENCES


Bishop, D. V. 2007 Using mismatch negativity to study central auditory processing in developmental language and literacy impairments: where are we, and where should we be going?. Psychol. Bull. 133, 651–672. (doi:10.1037/0033-2909.133.4.651)


Fiorentini, A. & Berardi, N. 1980 Perceptual learning specific for orientation and spatial frequency. Nature 287, 43–44. (doi:10.1038/287043a0)


Phil. Trans. R. Soc. B (2009)


Stein, J. & Walsh, V. 1997 To see but not to read; the magnocellular theory of dyslexia. Trends Neurosci. 20, 147–152. (doi:10.1016/S0166-2236(96)01005-3)


Phil. Trans. R. Soc. B (2009)


