Learning & Perception 1 (2009) 1, pp. 69–87
DOI: 10.1556/LP.1.2009.1.6

THE OTHER KIND OF PERCEPTUAL LEARNING

JÓZSEF FISER*

Department of Psychology and Volen Center for Complex Systems, Brandeis University,
Waltham, MA, USA

In the present review we discuss an extension of classical perceptual learning called the *observational learning* paradigm. We propose that studying the process how humans develop internal representation of their environment requires modifications of the original perceptual learning paradigm which lead to observational learning. We relate observational learning to other types of learning, mention some recent developments that enabled its emergence, and summarize the main empirical and modeling findings that observational learning studies obtained. We conclude by suggesting that observational learning studies have the potential of providing a unified framework to merge human statistical learning, chunk learning and rule learning.

Keywords: associative learning; rule learning; probabilistic models; chunking; Bayesian computation

1. INTRODUCTION

Perceptual learning has been traditionally defined as a practice-induced improvement in humans' ability to perform specific perceptual tasks. It "... encompasses parts of the learning process that are independent of conscious forms of learning and involve structural and/or functional changes in the primary sensory cortices" (Fahle and Poggio, 2002). In perceptual

* Address for correspondence; Volen Center for Complex Systems, Volen 208 MS 013, Brandeis University, Waltham, MA 02454, USA; Phone: (781)-736-3253; Fax: (781)-736-2398; E-mail: fiser@brandeis.edu

learning paradigms, the subject is presented with a well-defined task, such as orientation discrimination (Fiorentini and Berardi, 1980; Furmanski and Engel, 2000; Petrov, Dosher and Lu, 2006), texture discrimination (Ahissar and Hochstein, 1997; Karni and Sagi, 1991), motion direction discrimination (Matthews et al., 1999) or hyperacuity test (Poggio, Fahle and Edelman, 1992) that is explained verbally by the experimenter. After a repetitive training (typically including feedback), the subject's performance improves as quantified by threshold or reaction time measures. This experimental paradigm has been explored extensively both in the psychophysical (Dosher and Lu, 1998; Furmanski and Engel, 2000; Gold, Bennett and Sekuler, 1999) and neurophysiological domains (Gilbert, Sigman and Crist, 2001; Schoups et al., 2001), and the results obtained by such experiment define our views on all kind of learning that is positioned below high-level cognitive, categorical learning. In this article, I will make a case that, despite this widespread view, the classical paradigm and the kind of changes induced by it are only a subset, perhaps, not even the most important subset of the kind of perceptual learning we need to study in order to understand how the brain changes its internal representation to accommodate new aspects of the sensory input.

Arguably, research on learning aims at understanding the processes in the brain that allow humans and animals to adaptively react to their environment in the most efficient way. Focusing on the visual domain for now, this means to understand how humans and animals extract novel visual structures, features, or chunks from their visual environment and represent them so that later they can use these structures for the purposes of effective visual recognition and recognition-based action. This is the process by which infants make sense of their visual environment, learning the existence of objects and their interactions in scenes, and this is also the process by which adults acquire new visual knowledge of never before seen visual inputs. There are a number of essential characteristics of this process. First, it is implicit or unsupervised in that there is no explicit external direction as to what should or should not be retained from the visual input. Second, it works on an extremely complex, hierarchically structured, spatio-temporal input where not only are there many embedded substructures composing the scene, but many of these substructures participate in multiple contexts at different times. This requires a very sophisticated representation to be developed. Third, the interactions between objects in the outside world can alter the appearance of these structures substantially, thus their description must provide a large degree of generalization for successful application. Fourth, since a simple brute-force approach of learning all the possible features cannot cope with the complexity of real visual inputs due to the explosive combinatorics of the problem, the proposed learning mechanism must be computationally powerful enough to solve the task and meanwhile remain biologically plausible.

In this paper, I will argue that although the classically defined paradigm of visual perceptual learning investigates visual sensory learning, it is not the right paradigm to investigate how humans acquire higher-order visual structures in the natural environment. First, the task in structure learning is to automatically extract new descriptions from a large set of potential descriptions. In contrast, during classical perceptual learning subjects improve their threshold of discrimination between well-specified simple patterns, which can be done by increasing the sensitivity of existing detectors. Second, the result of structure-learning seldom if ever leads to conscious access of what has been learned. In contrast, during classical perceptual learning the attribute to be learned is pre-specified in the task description and, therefore, the learning process is much more influenced by cognitively controlled processes. These differences make it hard to see a direct link between the results obtained by the classic studies of perceptual learning and the goals of understanding how humans develop new visual representations.

But if not the classical paradigm then what is the appropriate approach to studying how humans do sensory learning that leads to useful representations? In the following sections, I will first specify a new paradigm of perceptual learning, called *observational learning paradigm* that appears to be a good candidate, and relate this paradigm to other known type of research areas of learning. Next, I quickly highlight some new developments that make it possible to make advances within the observational learning paradigm. After this, I present some experimental and modeling results within this paradigm and outline the new challenges that perceptual learning as a field faces in other to further our understanding about human representational learning.

2. A NEW KIND OF PERCEPTUAL LEARNING: OBSERVATIONAL LEARNING

I propose that there are three aspects of the classical perceptual learning that need to be altered for a new framework, the observational learning paradigm, to make it suitable for investigating how humans acquire new internal representations. The first is a conceptual one: although the overall goal of observational learning is to encode useful aspects of the low level sensory input for further processing, just like in the case of classical perceptual learning, the fundamental *computational* task is not that Gaussian noise obscures some of the elements of the input, and hence it prevents the correct perception of the input that could lead to learning of some useful underlying associations. Rather, the sensory input is fairly clear but ambiguous: it can support far too many possible combinations of elements that all could be potentially relevant higher-order features. This is a combinatorial problem and not a signal-processing problem that lays at the heart of classical perceptual learning studies. Thus, the new paradigm should be suitable to test this combinatorial aspect of the learning problem in a controllable way.

The second aspect is computational. As mentioned in the introduction, typical perceptual learning tasks use feedback, and thus they test how humans acquire new information in a supervised situation, while the principal challenge of developing new representations is to provide an unsupervised method of learning. Supervised learning is a much simpler task than unsupervised in the sense that there is a single well-defined objective function and, in each trial, there is a very detailed error information provided to re-tune the system. In contrast, the only objective function in unsupervised learning to capture the structure of the input. Even though it is clear that humans perform both unsupervised and supervised learning and there is a strong interaction between the two, the fundamental level of this process (translating light information into meaningful visual interpretations) is better characterized by unsupervised learning. While many processes of human knowledge acquisition are goal-directed and also rely on explicit external error-measures, the first step toward these more cognitive types of memory formations is a dimension-reducing unsupervised learning process. This process is based on the external structure of the visual world, its purpose is to develop a representation that support higher level learning, and it proceeds without a bias imposed by an explicit task. To investigate this process we need an experimental learning paradigm that is inherently unsupervised.

The third aspect is methodological. Since our goal is to investigate first the general mechanisms of visual learning we must use stimuli with characteristics that cannot contaminate the results because of some uncontrolled specific aspect of the stimuli. Simple visual stimuli can form intermediate-level features by recourse to already existing lower-level "grouping principles", and thus the internal representation on which learning occurs becomes inaccessible to the experimenter. For example, even though classical perceptual learning has been studied for decades, it is still not known what is learned exactly when the subject can perform the orientation discrimination task better. One solution to this problem is to have a total control over all statistics of the stimulus that the subject can utilize for learning. Paradoxically, this goal can best be achieved by not using the simplest stimuli. Even though displays with a number of localized Gabor-patches would be the type of fundamental stimuli traditionally used in visual psychophysics of perceptual learning, this is not the optimal stimuli for studying learning. Gabor-patches are very similar to each other, and thus when only a few Gabor-patches are used in each display, the visual input is too simple to provide the necessary environment for studying statistical learning. Conversely, when more than a few Gabor-patches are shown in the display, the visual system immediately invokes a number of previously developed mid-level representations based on configural or grouping mechanisms. For example, a cluster of semi-continuously oriented Gabor-patches will "pop out" of a background array of randomly oriented Gabor-patches (Kovacs et al., 1999). Thus, the statistical learning mechanism will work on these intermediate representations rather than the structure controlled explicitly by the experimenter. Thus the observational learning paradigm needs stimuli with completely controlled statistical features.

In the last ten years, we have developed an observational learning paradigm that fulfills the above requirements and is suitable for studying human representational learning both in adults and infants (Fiser and Aslin, 2001, 2002a, 2002b, 2005). We used arbitrary configurations of complex, highly discriminable novel shapes as the elements and generated visual scenes by combining a subset of these shapes in each scene according to some statistical rules. The general properties of statistical learning could be investigated using these complex shapes just as well as by using Gabor-patches or other simple stimuli but without interference from perceptual mechanisms already in place. Moreover, by randomly assigning shapes across subjects, we could eliminate any specific effect based on peculiar low-level feature co-occurrences, and control the available statistics very precisely. The paradigm is completely unsupervised and implicit, and it tests humans' ability of extracting higher order statistical regularities from unknown inputs. Thus, this observational learning paradigm allowed us to investigate the general statistical rules for how internal representations of new complex visual features naturally emerge.

3. RELATIONSHIP BETWEEN OBSERVATIONAL LEARNING AND OTHER TYPES OF LEARNING

Although I argued that observational learning is a necessary extension of perceptual learning that compensates for the insufficiency of classical perceptual learning in capturing the essential aspects of human learning, it is also necessary to relate observational learning to the other branches of learning to fully understand why this new paradigm is necessary.

One dimension along which different types of learning are ordered is the level of abstraction. Classical perceptual learning deals with the lowest level of changes focusing on basic characteristics of simple stimuli (such as orientation or position of a blob), while most other types of learning deal with either skill learning (e.g. typing), episodic or high level abstract category learning. The uniqueness of observational learning is that it links low level classical perceptual learning and high level abstract learning in a way that none of the traditional experimental paradigms do. For example, categorization has a subfield that is closely related to observational learning since it focuses on learning features that permit categorization (Smith and Medin, 1981). However, research on learning features for categorization, even for perceptual categories, typically focuses on the learning process within a specific task such as sorting out whether an input pattern belongs to categories A vs. B, and it is explicitly acknowledged that the identity of the task strongly influences the learning process (Ashby and Maddox, 2005). Thus, although categorization studies are related to observational learning, they typically do not address the automaticity and implicitness of this learning (but see Love, Medin and Gureckis, 2004).

The most fitting existing paradigm for observational learning is implicit learning, more specifically statistical learning. Implicit learning, broadly defined, is the ability to learn without awareness, and has been explored mostly in language acquisition and skill learning research (Stadler and Frensch, 1997). Although implicit learning is thought to be incidental but robust, there is no agreement in the field regarding the extent to which implicit learning produces abstract complex knowledge (Cleeremans, Destrebecqz and Boyer, 1998). Since early implicit learning studies were conducted in the domain of language acquisition, "abstract complex knowledge" has been defined initially as "knowing the grammar" that produces a set of observed strings (Reber, 1967). In implicit learning studies conducted in other domains, this concept of abstraction was generalized to mean "knowing the rules" that produced the observed inputs (Lewicki, Hill and Bizot, 1988). However, recent studies have proposed that using a grammar or a rule efficiently might not even require the explicit knowledge of the rules themselves (Perruchet and Pacton, 2006). It might be sufficient to use a simple associative learning mechanism that can trace distributional properties of the input (by statistical learning) or a chunking mechanism that finds small typical fragments of the input (by chunk learning) to account for all the rule-learning results of implicit learning experiments.

Statistical learning is a special version of implicit learning that instead of symbolic rules focuses on humans' ability to learn the simple statistical structures of the input. Statistical learning was first also introduced in the domain of language acquisition (Saffran, Aslin and Newport, 1996; Saffran et al., 1997). These studies showed that mere exposure to a continuous sequence of auditory syllables, visual shapes or full visual scenes is sufficient for adults and infants to extract regularities from the input. Specifically, in the Saffran et al. studies subjects incidentally learned the transitional probabilities between syllables presented in a continuous uninterrupted auditory stream generated by a certain rule, that is, they automatically learned which syllable was most likely to follow a given syllable. Statistical learning was extended to the domain of touch (Conway and Christiansen, 2006; Hunt and Aslin, 2001) and even to other species (Hauser, Newport and Aslin, 2001) suggesting that it is a very general domain-independent behavioral phenomenon. Since statistical learning is unsupervised and implicit is uniquely adequate for observational learning both in adults and infants. Indeed, this is the paradigm that we modified in the modality of vision to conduct perceptual learning studies in an observational learning framework (Fiser and Aslin, 2001, 2002a, 2002b). Even though these experiments used composition of simple black shapes as stimuli, they clarified the computational basis of human feature learning and thus they opened the road for non-classical perceptual learning experiments with more realistic stimuli that can directly target the issue of what humans' internal visual representation of the external world is and how it develops.

4. NEW DEVELOPMENTS FOSTERING THE EMERGENCE OF THE OBSERVATIONAL LEARNING PARADIGM

Recently, there were significant theoretical, conceptual, and technological advancement that allowed us to develop a Bayesian computational framework of the observational learning paradigm. On the theoretical level, a new development in the mathematics of probability expanded the scope and explanatory power of probabilistic models (Jordan, 1998), thus the new Bayesian methods can learn far more complex tasks than the standard neural network learning algorithms introduced in the connectionist framework two decades ago (Rumelhart, McClelland and PDP Research Group, 1986). On the conceptual level, in cognitive psychology and in computational neuroscience, the traditional deterministic view on brain functioning and behavior has been more and more strongly challenged by a probabilistic view following Helmholtz's original idea (Helmholtz, 1925). According to this view, the brain can be viewed as making "unconscious inferences" to predict likely outcomes in the face of inherent uncertainty of the situation due to insufficient data (Knill and Pouget, 2004). Based on these developments, a large number of successful models have been published in the domains of sensory psychophysics (Kersten, Mamassian and Yuille, 2004), sensorimotor integration (Kording and Wolpert, 2006), higher cognitive processes (Tenenbaum, Griffiths and Kemp, 2006) as well as in classical conditioning (Courville et al., 2003; Courville, Daw and Touretzky, 2004) and modeling neural coding in the brain (Ma et al., 2006). The natural framework of these models is a Bayesian formalism, and since in recent years many earlier connectionist learning models were also reformulated in these terms, the Bayesian framework provides the most comprehensive approach for investigating human learning.

In the domain of technology, both multi-electrode recordings from awake animals and imaging studies radically altered our view of how the brain works (Buzsaki, 2004; Buzsaki and Draguhn, 2004; Fiser, Chiu and Weliky, 2004; Fox and Raichle, 2007; Gusnard and Raichle, 2001; Kenet et al., 2003). These studies provided glimpse of large-scale neural activity in the nervous system and they demonstrated the existence of widespread spontaneous activity in all areas of the brain. Given that this spontaneous activity is not random but highly organized (Arieli et al., 1995; Fiser, Chiu and Weliky, 2004; Tsodyks et al., 1999) and that neural activity in the brain is very expensive metabolically speaking (Attwell and Laughlin, 2001; Lennie, 2003), a number of researchers suggested that spontaneous activity might have a functional role rather than being noise in the nervous system (Fiser, Chiu and Weliky, 2004; Harris, 2005). Clearly, spontaneous activity cannot be in direct connection with instantaneous visual input, and thus it must represent other kind of information related to the internal state of the system, which can encompass goals, stimuli experienced earlier, and additional knowledge of the observer. The issue of how this kind of information can be combined with information about momentarily perceived input is naturally linked to probabilistic frameworks exemplified by the Bayesian approach.

These developments promote a new link between psychological and physiological studies of learning. Learning is an enormous and complex field of research in neuroscience, but most of its effort is focused on understanding the underlying molecular and cellular processes of experience induced changes in the nervous system (Martinez and Kesner, 2007). Much fewer studies concentrate on explaining learning on system level, and there are very few comprehensive theories of learning on what Marr called the “computational level” of the problem (Marr, 1982). The probabilistic approach offers such a theory by investigating the issue of perceptual learning from a normative standpoint, and the observational learning paradigm is tailored to provide an easy way to connect physiological and psychological results of learning.

5. RESULTS OF VISUAL OBSERVATIONAL LEARNING

In this section, I review some of the most important findings we obtained by using the observational learning paradigm. We used the paradigm to explore the question posed in the introduction: What are the novel visual structures, features, or chunks humans extract from newly encountered visual inputs, and how do they represent them so that later they can use these structures for the purposes of effective visual recognition? The results demonstrate not only the advantages of this paradigm, but also the way these investigations can be tied to computational models to gain further insights of the nature of this kind of perceptual learning.

5.1. Humans are automatically sensitive of “suspicious coincidences” from birth

All the subsequent visual learning experiments followed the same basic design. We use (typically twelve) black shapes on a white background and create an inventory of structured building blocks called combos. A combo is a set of two or more shapes in a particular spatial (or temporal) configuration: any time one element of the combo appears, all other elements of the combo will also appear in the predefined spatial (or temporal) configuration. The inventory of combos is then used for generating a large number of scenes by selecting a number of combos randomly for each scene and placing them on a grid in various adjacent configurations so that none of the combos is completely separated from the other combos in the scene *(Fig. 1)*. The

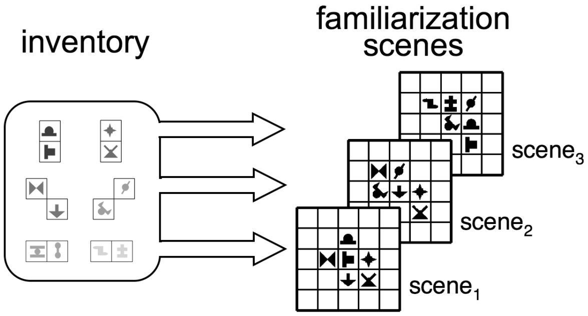

Figure 1. Generating familiarization scenes

Left: Twelve shapes organized into six pair combos of particular spatial arrangement constitutes the inventory (colors are used for demonstration purposes only). Right: Sample scenes generated from the combos using three randomly selected combos at a time and arranging them randomly on a 5 × 5 grid.

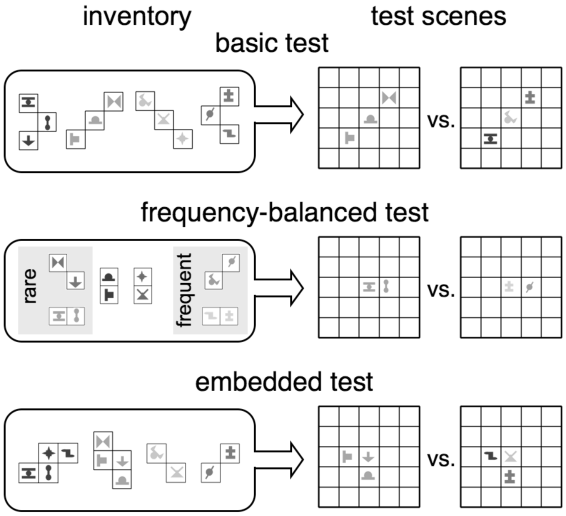

Figure 2. Three types of test used in the visual learning experiments

Left: a typical inventory of the given test type. Right: a typical example of a test trial. All gray shaded are for demonstration only, in the experiment only black color was used. Top: A true triplet combo compared to a random triplet. All individual shapes appeared equal number of times during practice. Middle: A true pair combo compared to a pair of elements selected from frequent combos. The elements of the two pairs co-occurred equal number of times, but elements of the true combo always appeared together whereas elements of the pair in the right could equally often appear without each other. Bottom: A part of a true quad combo compared to a random triplet. All individual shapes appeared equal number of times during practice.

resulting scenes contain six or more shapes shown on a rectangular grid so that there are no grouping cues for shape combos except the statistical co-occurrence of shapes across scenes.

An experiment is then conducted in two phases. In the first familiarization phase, subjects see a long sequence of multi-shape scenes, each two seconds long with a one second pause between them. The subjects have no explicit task other than paying attention to the scenes and they receive no feedback of any kind. In the second testing phase, a set of 2-AFC trials are given to the subject. In each trial two displays are presented for one second, one of which shows a single combo or an embedded part of a combo, and the other shows randomly selected individual shapes arranged in combo format. The subject has to choose the arrangement that looks more familiar based on the prior familiarization scenes. Notice, that subjects have not seen any of the combos alone during familiarization, thus both test displays were novel. Each subject's preference for the combo or combo-part over the random shape combination is taken as evidence that the subject extracted (unconsciously and automatically) the underlying structure of the scenes.

Using this experimental design we have run a large number of visual statistical learning experiments in the last couple of years. The goal of the first set of experiments was to establish whether humans are sensitive to second order joint and conditional visual statistics in the spatial and temporal domains, if yes then how universal this sensitivity is across adults and infants. The significance of this question come from the fact that it has been established a long time ago that efficient learning of internal representations of highly complex input is possible only if humans are capable of extracting such statistics (Barlow, 1989, 1990). Nevertheless, whether humans are truly capable to do this has never been tested empirically.

First, we ran the simplest spatial version of the experiment with adults using the inventory of six combo pairs (144 different 6-shape scenes) and testing for pairs of shapes (Fiser and Aslin, 2001). We found that after 6–8 minutes of continuous exposure to the familiarizations scenes, adults became sensitive to both the co-occurrence of shapes within the combos and the correlation between shapes and their position on the grid *(Fig. 3)*. In Experiment 1, subjects could rely on both shape co-occurrences and shape-position associations. In Experiment 2 only the number of shape co-occurrences differed between the true combo and the random pair. Both tests showed a significant preference to true combos and there was a significant difference between the results of the two experiments.

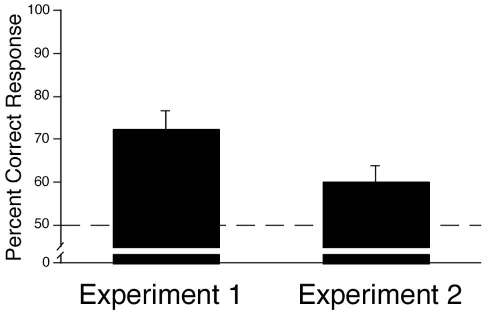

Figure 3. The results of the first two visual statistical learning experiments. Bars show percent of responses selecting the true pair combos, error bars show SEM, 50% is chance.

In a third experiment we used different appearance frequencies for different combos during training in order to create a situation where the co-occurrence of shapes within some combos was equated with the accidental co-occurrence of two shapes that did not belong to the same combo. With this *frequency-balanced* test we could show that humans preferentially encode combos with two elements perfectly predicting each others' appearance compared to a pair of shapes that appeared together the same number of times as the true combo but which could also appear without each other *(Fig. 4)*. In other words, humans were sensitive to the conditional probability statistic of element pairs even when the joint probabilities were equated. Sensitivity to conditional probabilities in the spatial modality amounts to sensitivity to transitional probabilities in the temporal modality, which has been demonstrated in the auditory domain before (Aslin, Saffran and Newport, 1998), and we have also demonstrated it in a series of experiments in the visual domain (Fiser and Aslin, 2002a). In the third experiment, we also found that apart from the conditional probabilities between shapes, humans maintain sensitivity to the frequency of the individual shapes, since they could readily select the shapes that appeared more often during the practice session *(Fig. 4)*.

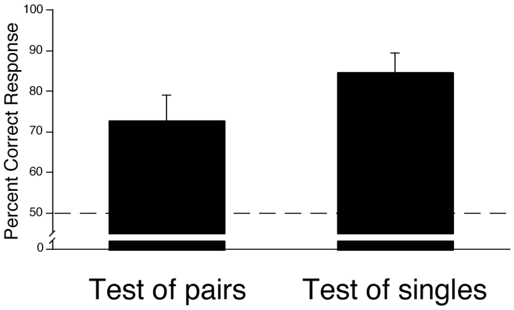

Figure 4. The results of the frequency-balanced experiment. Bars show percent of responses selecting the true pair combos, error bars show SEM, 50% is chance performance.

To determine whether this ability to extract statistics incidentally from visual scenes is a fundamental learning tool for humans at all ages, we tested whether these sensitivities are present in infants as well. Simplified versions of the above experiments were run on 8-month-old infants in a preferential looking paradigm (Fiser and Aslin, 2002b). Here the combos were pairs of elements, but the scenes consisted of only three elements – a combo and a corresponding noise element that could appear in arbitrary position around the combo. We found that after about 2 minutes of familiarization with the scenes, infants looked significantly longer at the picture of a base-pair (a combo of the inventory) than at a randomly combined pair of shapes from two combos *(Fig. 5)*. This replicated the results of the first two adults experiments.

We then asked whether infants would also perform well (as adults did) with frequency-balanced pairs. Infants looked significantly longer at the original combos than at the frequency-balanced random shape pairs *(Fig. 6)*. Interestingly, we found that infants did not replicate the adult results from the single shape test: they were unable to distinguish between low frequency and high frequency shapes. Together, these results confirmed that human adults and infants have access to information about conditional probability relations in previously unknown visual scenes and thus in principle it is possible for them to extract higher order visual structures by statistical learning mechanisms.

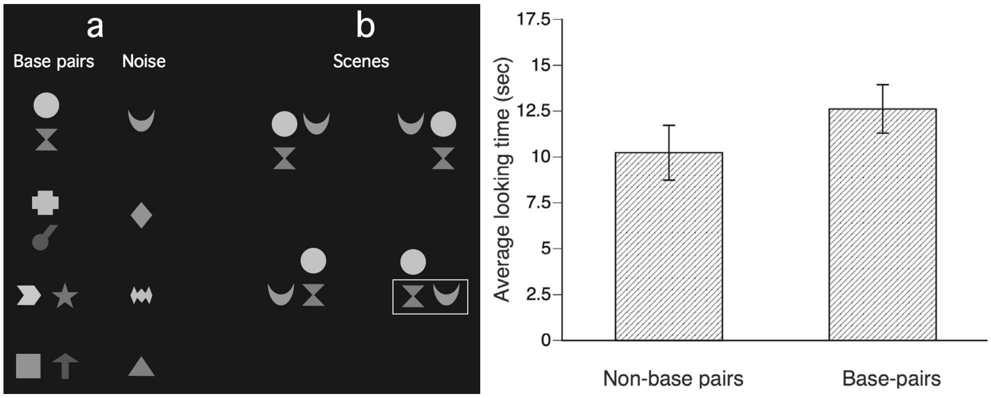

Figure 5. The infant visual statistical learning study

Left: a) Inventory of the experiments with four combos (base-pairs) and four noise elements. b) Four possible scenes composed from one combo and its noise element. The white rectangular indicates the frequency balanced pair used in the second pair test. Right: Results of the first pair test. Error bars show SEM. Babies looked significantly longer at true combos.

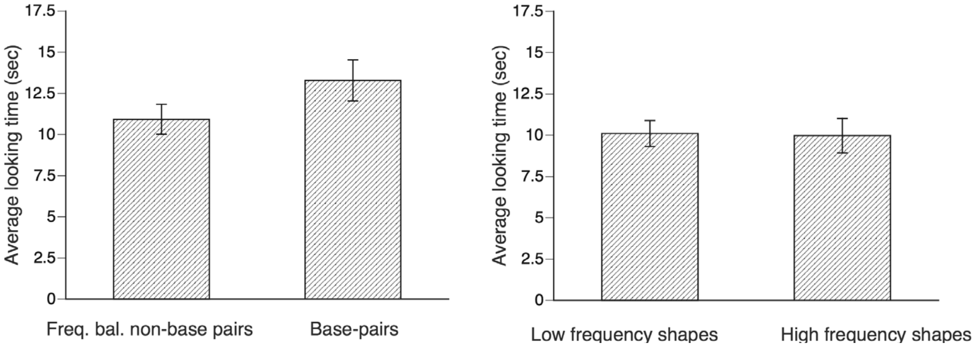

Figure 6. Results of the frequency-balanced infant learning experiment

Left: Results of the pair test. Babies looked longer at rare true combos. Right: Results of the single shape frequency test. There was no reliable difference in looking times between frequent and rare elements. Error bars show SEM.

5.2. Humans encode visual information in a minimally sufficient manner

In the next set of experiments, we turned to the well-known problem of the "curse of dimensionality" or "combinatorial explosion" of statistical learning (Bellman, 1961; von der Malsburg, 1995). Briefly, this problem comes from the fact that statistical learning of higher-order structures in a complex environment is impossible even theoretically because the number of training examples needed for learning the right structures is prohibitively large. This hard limit on brute force learning is in direct contradiction with the claim that humans learn their internal visual representations by statistical learning. In order to investigate this issue, we used visual stimuli with hierarchical internal structure, where the number of available statistics for learning the underlying structure of the scene grew exponentially with the number of shapes involved in defining the underlying combos of the inventory (Fiser and Aslin, 2005). In the first experiment, we used the inventory of four triplet combos and tested whether humans became sensitive to both the triplet structure and the structure of the pairs embedded in the triplet. We found that adults reliably selected the combo triplets over random shape triplets, but there was no such preference for the embedded pairs over random pairs (*Fig. 7,* left).

In the next experiment, we increased the size of the combos, and used the inventory of two quadruples and two pairs to construct the scenes. During test, subjects had to chose between a quad combo and a random quad, a pair combo and random pairs, and an embedded pair and random pair. Subjects preferred the quad and pair combos to the random structures, but they were at chance performance when they chose between a shape pair embedded in a quad and a random pair (*Fig. 7,* right).

This pattern did not change when we doubled the familiarization time (Fig. 7, right, shaded bars). Interestingly, in a subsequent study we found that embedded triplets of the same inventory were preferred over random triplets. In a series of control experiments, we ruled out a number of alternative explanations, and concluded that this pattern of results suggests that subject generate a minimally sufficient representation of combos instead of encoding the full structure of the underlying scenes (Fiser and Aslin, 2005a). We also pointed out that such a strategy could avoid the curse of dimensionality and naturally fits into a Bayesian framework.

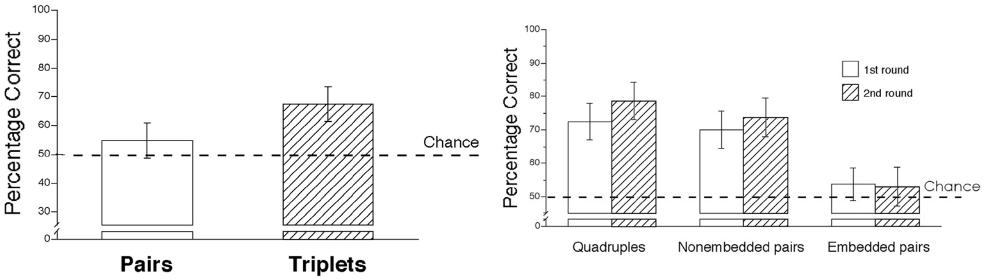

Figure 7. Results of the two embedded learning experiments

Left: Results of the embedded experiment with the inventory of triplet combos. Right: Results of the embedded experiment with the inventory of quadruple and pair combos. In both experiments, subjects reliably learned the true combos.

5.3. Humans learn structure of their visual environment in a statistically optimal way

Using a Bayesian framework, we have developed a computational model to explain all of our learning results, and we compared this model to other existing models of human unsupervised learning (Orbán et al., 2008). The previously proposed models can be grouped into roughly four types. Models of type-1 are based on frequency counting, where the learning algorithm keeps track of the occurrence of individual patterns or events and patterns with the highest frequencies represent the significant memory traces. This is the simplest type of learning model and counting episodic memories belongs to this class. Type-2 models keep counts of pairwise co–occurrences of elements in the scenes. Apart from the important issue of how to define what constitutes an element, this learning model is superior to type-1 models because instead of a single holistic representation it operates on a vocabulary-based combinatorial description. Type-3 models compute conditional or transitional probabilities between two elements of the scene rather than just their co-occurrence frequencies, and stores combinations with high conditional probability as useful memory traces. Features obtained by this kind of learning are the "spurious coincidences" in the input stimulus explained by Barlow (Barlow, 1989, 1990, 2001). Type-3 models of learning are the most popular among empirical researchers of statistical learning (Kuhl, 2004). In contrast to the first three models, type-4 models are not simple event counters, but rather full probabilistic models that learn all pair-wise correlations between elements across all the scenes. The crucial difference between models of types-3 and -4 is that type-4 models assess the familiarity or likelihood of a new input based on not only the elements that are present but also on elements that are absent in the input. A type-4 model is a statistically optimal implementation of the most extensively studied associative learning mechanisms in the literature including feature learning mechanisms that are typically modeled in unsupervised neural network architectures (Dayan and Abbott, 2001).

The model that we developed to test against these previous models, which we refer to as *"ideal learner"*, is an explicit chunking model and computationally it is based on Bayesian model averaging. Bayesian model averaging is a statistically principled optimal way of solving the following problem: given a set of atomic elements, how to select an inventory of chunks based on those elements that allows capturing the input in a minimally sufficient way while it also allows for a maximum ability of generalizing to inputs never experienced before. In our model, each selected inventory (each choice of what combos/chunks will be the building blocks of scenes) defines a probability distribution over all possible scenes: P(scene$_1$, scene$_2$, …, scene$_n$| Inventory). Chunks were formalized as latent variables that describe the identity and relative position of shapes making up the chunk. If the chunk is present in a scene, the probability of these shapes in the given configuration increases, if the chunk is absent, each element of the chunk can still appear with a "spontaneous" probability regardless what the other shapes do. Thus observing the scenes, chunks can be inferred as spurious coincidences of shapes, and for any set of chunks defining an inventory, the probability of familiarization scenes can be computed. Based on these probabilities and assuming that chunks appear independently from each other, the best inventory can be selected by Bayesian model averaging by ranking the inventories according to the summed probability assigned by the inventory to the set of practice trials. Due to the "automatic Occam's razor" effect of Bayesian model averaging (MacKay, 2003), this method will select the optimal inventory that describes the previous practice scenes sufficiently well but does not prevent generalization to novel scenes (Orbán et al., 2008).

We tested which of these five models best captures human performance in the learning experiments described above (Orbán et al., 2008). For this, we imported the training and test stimuli of each experiment, and trained and tested each model exactly the same way as we did with humans. *Figure 8* summarizes our findings. In the simplest test (panel a), the inventory contained six equal-frequency pairs. Only type-1 failed to predict above-chance human performance on the basic test of true pairs vs. mixture pairs. In the frequency balanced test (panel b), the inventory contained six pairs of varying frequency, and both type-1 and type-2 failed to predict above-chance human performance on the test of true rare pairs vs. frequency-balanced mixture pairs. In the simple embedded experiment, (panel c), the inventory contained four equal-frequency triplets, and human performance was above chance on the basic test of true triplets vs. mixture triplets and at chance on the test of embedded pairs vs. mixture pairs. Types-1-2-3 incorrectly predicted the same performance on the basic and embedded tests. Finally, in the second embedded experiment (panel d), the inventory contained two quadruples and two pairs, all with equal frequency. Human performance was above chance on the basic tests of true quadruples or pairs vs. mixture quadruples or pairs, and on the test of embedded triplets vs. mixture triplets, but it was at chance on the test of embedded pairs vs. mixture pairs. Types-1-2-3 incorrectly predicted the same performance on all tests. Only type-4 and the ideal Bayesian learner captured the overall pattern of human performance in all these ex-

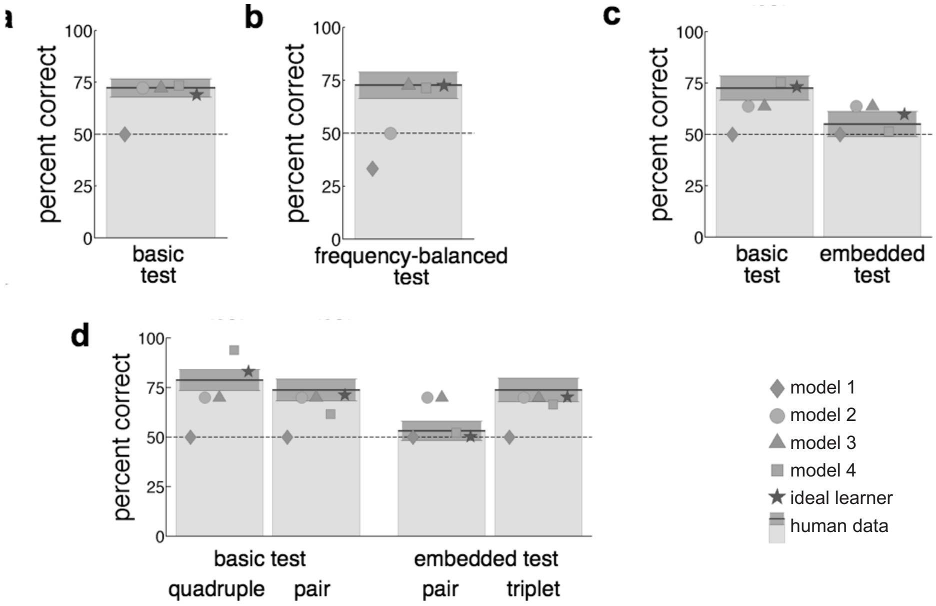

Figure 8. Summary of human performance (gray bars) and model predictions (symbols) for a series of experiments from (Fiser and Aslin, 2001, 2005) using increasingly complex inventories. The black line and dark gray region at the top of each bar indicates mean and SEM, respectively.

periments. In summary, we found that only the ideal learner could qualitatively and quantitatively replicate all the patterns of human performance in all experiments.

However, the type-4 model, which is most widely considered the appropriate model of human learning, performed relatively well and based on these results it cannot be discarded as a good model of human learning. Since type-4 models and the ideal learner follow a very different computational philosophy, however, we could design an experiment in which the prediction of the two models is contradictory and compare their results to human performance. Specifically, the inventory of the experiment had four so-called "circular triplets", four single elements that also formed sometimes a quadruple and two pairs *(Fig. 9)*. The key element of the design was that the four triplets shared shapes and the construction of the scenes was such that assessing only pair-wise correlations could not give any help in establishing the identity of the four circular triplets. Similarly, the four singletons appeared many times alone but also together as a quadruple so that their pair-wise correlations were also balanced. Thus from the standpoint of first- and second-order correlations, there was no difference between the circular triplets and triplets composed from the singletons of the inventory. However considering higher order correlations, the circular triplets were true building blocks of the inventory, whereas the singleton-based triplets were just faulty associations that did not capture the significant chunks in the scenes precisely enough.

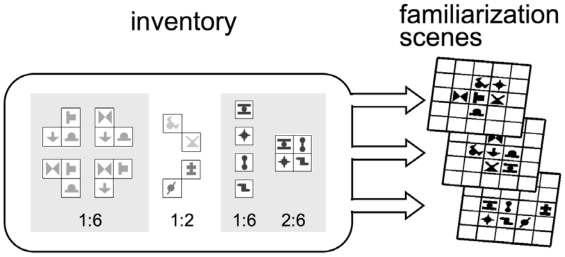

Figure 9. The inventory of the experiment testing the difference between type-4 models and the ideal learner. Combos in the left and right shaded areas represent the circular triplets and the singleton combos, respectively. Ratios at the bottom show the relative frequency of the combos across the practice scenes. All element frequencies appearance and pair co-occurrence frequencies were equated.

We compared the prediction of the models with human performance in three tests: circular triplets vs. random triplets, singleton-based triplets vs. random triplets, and circular vs. singleton-based triplets *(Fig. 10)*. Humans showed a significant preference for circular triplets in the first and third tests and a chance performance in the second test. The pattern of predictions showed a clear difference between model 4 and the ideal learner. We used the models with their parameter tuned to account the results by the previous experiments to predict the model performances in the present experiment. The ideal learner followed human performance in all three experiments. In the first one, it correctly showed that circular triplets are true building blocks of the visual scenes. In the second test, it correctly signaled, that a singleton-based triplet is not a very significant chunk of the visual scenes despite the fact that that triplet configuration appeared twice as often across the set of scenes as a circular triplet did. In the third, di-

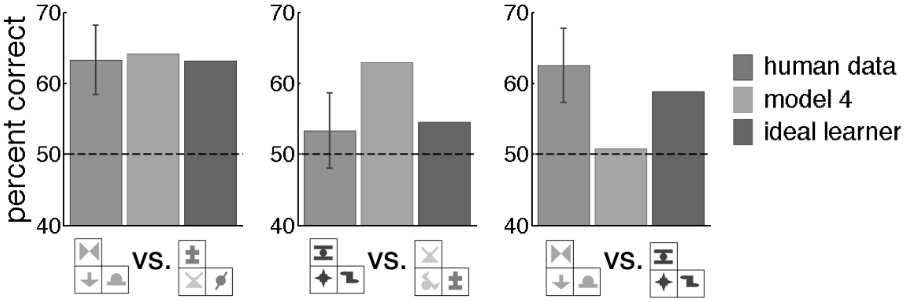

Figure 10. The results of the experiment testing the difference between type-4 models and the ideal learner. Left panel show the comparison between circular triplets and mixed triplets, the middle panel shows singleton-based triplets versus mixed triplets and the right panel shows the direct comparison of circular triplets to singleton-based triplets. Only the ideal learner followed the pattern of human subjects' test performance.

rect comparison, it favored significantly the circular triplets over singleton-based triplets. In stark contrast, the predictions of model 4 were in direct contradiction with human performance by completely equating the significance the two types of test triplets. As a result, even though in the first test the model correctly favored the circular triplets over random ones, it favored equally the singleton-based triplets in the second test, and it could not make a difference between the two in the third direct test either. These results suggest that rather than learning all pair-wise correlations of the input, humans follow an optimal model averaging aiming at a minimally sufficient chunk-based representation of the input.

6. CONCLUSIONS

The above examples highlight the distinct advantages of investigating human learning within an observational learning paradigm. First, the methodology can be easily transfer from adult to infant studies and even to animal models, which is important if the developmental and neural aspects of the process need to be studied. Second, the computational problem of extracting statistically significant structures of the input is naturally linked to the main issues of vision, namely to feature extraction, perceptual grouping, recognition and categorization. Third, placing the problem in a clear statistical framework helps clarifying the true nature of the computations performed by the visual system which leads to deeper understanding of the problem such as whether associative learning is sufficient to capture human learning.

Although the present results represent a significant advancement in our understanding of human visual learning, observational learning is far away from being completely set up for asking every pertinent question about how humans acquire their internal visual representations. The present implementation of the paradigm focused on the purest measurement of statistical dependencies by working with highly abstracted stimulus set and therefore excluding the opportunity to use readily available prior knowledge about the visual environment. Now that the basic principles of learning are clarified, the next step is exploring what this prior knowledge is and how it is integrated with the bottom-up information when a stimulus is presented. This can be explored in two ways, first by still using abstract stimuli but measuring the available prior knowledge for the system and see how this knowledge influences visual processing and learning. Second, by changing the stimulus set so that the information provided about its structure would be carried by the low-level dimensions of natural vision such as contrast, orientation, spatial frequency, or color. Using these methods, the primary challenge for the observational learning paradigm is to find computational and experimental evidence that the mechanisms responsible for statistical learning of simple spatial and temporal structures underlie a general process that can capture human visual learning from simple features to complex visual structures under natural conditions. If it succeeds, this could expand the notion of perceptual learning to encompass not only statistical learning but also chunk learning and rule learning under a unified computational principle.

ACKNOWLEDGEMENTS

The various experimental and modeling results described in the present paper have been obtained through collaborative effors with Richard Aslin, Gergő Orbán and Máté Lengyel.

REFERENCES

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

Arieli, A., Shoham, D., Hildesheim, R., Grinvald, A. (1995): Coherent spatiotemporal patterns of ongoing activity revealed by real-time optical imaging coupled with single-unit recording in the cat visual cortex. *Journal of Neurophysiology, 73,* 2072–2093.

Ashby, E. G., Maddox, W. T. (2005): Human category learning. *Annual Review of Psychology, 56,* 149–178.

Aslin, R. N., Saffran, J. R., Newport, E. L. (1998): Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*(4), 321–324.

Attwell, D., Laughlin, S. B. (2001): An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism, 21,* 1133–1145.

Barlow, H. B. (1989): Unsupervised learning. *Neural Computation, 1,* 295–311.

Barlow, H. B. (1990): Condition for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research, 30,* 1561–1571.

Bellman, R. (1961): *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press.

Buzsaki, G. (2004): Large-scale recording of neuronal ensembles. *Nature Neuroscience, 7*(5), 446–451.

Buzsaki, G., Draguhn, A. (2004): Neuronal oscillations in cortical networks. *Science, 304*(5679), 1926–1929.

Cleeremans, A., Destrebecqz, A., Boyer, M. (1998): Implicit learning: News from the front. *Trends in Cognitive Sciences, 2*(10), 406–416.

Conway, C. M., Christiansen, M. H. (2006): Statistical learning within and between modalities – Pitting abstract against stimulus-specific representations. *Psychological Science, 17*(10), 905–912.

Courville, A. C., Daw, N. D., Gordon, G. J., Touretzky, D. S. (2003): Model uncertainty in classical conditioning. In *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Courville, A. C., Daw, N. D., Touretzky, D. S. (2004): Similarity and discrimination in classical conditioning: A latent variable account. In *Advances in Neural Information Processing Systems.* Cambridge, MA: MIT Press.

Dayan, P., Abbott, L. F. (2001): *Theoretical Neuroscience*. Cambridge, MA: MIT Press.

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.

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

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

Fiser, J., Aslin, R. N. (2001): Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499–504.

Fiser, J., Aslin, R. N. (2002a): Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology-Learning Memory and Cognition, 28,* 458–467.

Fiser, J., Aslin, R. N. (2002b): Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences USA, 99,* 15822–15826.

Fiser, J., Aslin, R. N. (2005): Encoding multielement scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General, 134,* 521–537.

Fiser, J., Chiu, C. Y., Weliky, M. (2004): Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature, 431,* 573–578.

Fox, M. D., Raichle, M. E. (2007): Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience, 8*(9), 700–711.

Furmanski, C. S., Engel, S. A. (2000): Perceptual learning in object recognition: object specificity and size invariance. *Vision Research, 40,* 473–484.

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

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

Gusnard, D. A., Raichle, M. E. (2001): Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience, 2,* 685–694.

Harris, K. D. (2005): Neural signatures of cell assembly organization. *Nature Reviews Neuroscience, 6*(5), 399–407.

Hauser, M. D., Newport, E. L., Aslin, R. N. (2001): Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78*(3), B53–B64.

Helmholtz, H. v. (1925): *Treatise on Physiological Optics*. Translated from the 3rd German edition (1910), Southall, J. P. C. (ed.). Washington, D. C.: Optical Society of America.

Hunt, R., Aslin, R. N. (2001): Statistical learning in a serial reaction time task: Simultaneous extraction of multiple statistics. *Journal of Experimental Psychology: General, 130,* 685–680.

Jordan, M. I. (ed.) (1998): *Learning in Graphical Models*. Dordecht, The Netherlands: Kluwer Academic Publisher.

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.

Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A., Arieli, A. (2003): Spontaneously emerging cortical representations of visual attributes. *Nature, 425,* 954–956.

Kersten, D., Mamassian, P., Yuille, A. (2004): Object perception as Bayesian inference. *Annual Review of Psychology, 55,* 271–304.

Knill, D. C., Pouget, A. (2004): The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719.

Kording, K. P., Wolpert, D. M. (2006): Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10*(7), 319–326.

Kovacs, I., Kozma, P., Feher, A., Benedek, G. (1999): Late maturation of visual spatial integration in humans. *Proceedings of the National Academy of Sciences USA, 96*(21), 12204–12209.

Kuhl, P. K. (2004): Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience, 5,* 831–843.

Lennie, P. (2003): The cost of cortical computation. *Current Biology, 13,* 493–497.

Lewicki, P., Hill, T., Bizot, E. (1988): Acquisition of procedural knowledge about a pattern of stimuli that cannot be articulated. *Cognitive Psychology, 20*(1), 24-37.

Love, B. C., Medin, D. L., Gureckis, T. M. (2004): SUSTAIN: A network model of category learning. *Psychological Review, 111*(2), 309–332.

Ma, W. J., Beck, J. M., Latham, P. E., Pouget, A. (2006): Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9*(11), 1432–1438.

MacKay, D. J. C. (2003): *Information Theory, Inference, and Learning Algorithms.* Cambridge, UK: Cambridge University Press.

Marr, D. (1982): *Vision*. San Francisco: W.H. Freeman.

Martinez, J., Kesner, R. (eds) (2007): *Neurobiology and Learning and Memory* (2nd ed.). Burlington, MA: Academic Press.

Matthews, N., Liu, Z., Geesaman, B. J., Qian, N. (1999): Perceptual learning on orientation and direction discrimination. *Vision Research, 39,* 3692–3701.

Orbán, G., Fiser, J., Aslin, R. A., Lengyel, M. (2008): Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences USA, 105,* 2745–2750.

Perruchet, P., Pacton, S. (2006): Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10*(5), 233–238.

Petrov, A. A., Dosher, B. A., Lu, Z. L. (2006): Perceptual learning without feedback in non-stationary contexts: Data and model. *Vision Research, 46*(19), 3177–3197.

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

Reber, A. S. (1967): Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*(6), 855ff.

Rumelhart, D. E., McClelland, J. L., PDP Research Group (eds) (1986): *Parallel Distributed Processing – Explorations in the Microstructure of Cognition* (Vols 1–3). Cambridge, MA: The MIT Press.

Saffran, J. R., Aslin, R. N., Newport, E. L. (1996): Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., Baruecco, S. (1997): Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*(2), 101–105.

Schoups, A., Vogels, R., Qian, N., Orban, G. (2001): Practising orientation identification improves orientation coding in V1 neurons. *Nature, 412*(6846), 549–553.

Smith, E. E., Medin, D. L. (1981): *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Stadler, M. A., Frensch, P. A. (eds) (1997): *Handbook of Implicit Learning*. Thousand Oaks, CA: Sage Publications, Inc.

Tenenbaum, J. B., Griffiths, T. L., Kemp, C. (2006): Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309–318.

Tsodyks, M., Kenet, T., Grinvald, A., Arieli, A. (1999): Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science, 286,* 1943–1946.

von der Malsburg, C. (1995): Binding in models of perception and brain function. *Current Opinion in Neurobiology, 5,* 520–526.