Repetition and the brain: neural models of stimulus-specific effects

Kalanit Grill-Spector\textsuperscript{1,2}, Richard Henson\textsuperscript{3} and Alex Martin\textsuperscript{4}

\textsuperscript{1}Department of Psychology, Stanford University, Stanford, CA 94305, USA
\textsuperscript{2}Neurosciences Program, Stanford University, Stanford, CA 94305, USA
\textsuperscript{3}MRC Cognition and Brain Sciences Unit, Cambridge, UK
\textsuperscript{4}Laboratory of Brain and Cognition; National Institute of Mental Health, Bethesda, Maryland 20892-1366, USA

One of the most robust experience-related cortical dynamics is reduced neural activity when stimuli are repeated. This reduction has been linked to performance improvements due to repetition and also used to probe functional characteristics of neural populations. However, the underlying neural mechanisms are as yet unknown. Here, we consider three models that have been proposed to account for repetition-related reductions in neural activity, and evaluate them in terms of their ability to account for the main properties of this phenomenon as measured with single-cell recordings and neuroimaging techniques. We also discuss future directions for distinguishing between these models, which will be important for understanding the neural consequences of repetition and for interpreting repetition-related effects in neuroimaging data.

Introduction

When stimuli are repeated, neural activity is usually reduced. This neural repetition effect has been reported at multiple spatial scales, from the level of individual cortical neurons in monkeys [1–3] to the level of hemodynamic changes (measuring the pooled activation of millions of neurons) in humans using functional magnetic resonance imaging (e.g. fMRI [4–10]). Repetition-related reductions also occur at multiple temporal scales, both in their longevity – from milliseconds [3] to minutes [9] and days [11] – and in the latency of their expression [12,13]. The phenomenon also occurs in multiple brain regions, and across an impressively large number of experimental conditions.

This stimulus-specific reduction in neural activity has been referred to as adaptation [14–16], mnemonic filtering [17], repetition suppression [18], decremental responses [19] and neural priming [20]. We will use ‘repetition suppression’ (RS) to refer to decreased neural responses following stimulus repetition. As will be apparent below, it is important to distinguish the scale at which RS arises, because the correspondence across scales (e.g. neural firing and hemodynamic responses) might not be simple.

Interest in repetition effects has recently intensified, for two main reasons. First, repetition effects have proved useful for inferring the nature of representations across different stages of a processing stream. This approach has been used behaviorally (e.g. using visual aftereffects to infer the nature of orientation tuning [21] or face representation [22,23]), with single-cell recording [24], and more recently has become popular with fMRI, particularly given claims that it enables improved spatial resolution [7] (Box 1). The second reason for heightened interest is the possibility that RS might be the neural correlate of priming [25–27]. Priming refers to improved processing of a repeated stimulus according to some behavioral measure (e.g. greater accuracy in identifying the stimulus, or faster response times to make a decision about it), and often occurs under the same experimental conditions as RS. Nevertheless, it is important to note that, under certain conditions, priming can be associated with increased activity, rather than reduction (for discussion of repetition enhancement effects and changes in frontal cortices that might contribute to priming, see [27,28]).

The purpose of this review is to consider several kinds of neural models that have been proposed to account for repetition suppression (RS). We focus primarily on studies using visually presented objects and their effects on the ventral object processing stream, to maximize overlap between monkey and human studies. We evaluate the neural models in terms of their ability to account for the main properties of RS as measured with single-cell recordings, fMRI and electroencephalogram/magnetoencephalogram (EEG/MEG), discuss implications of these models for interpreting experimental results, and propose directions for distinguishing between the models.

Repetition suppression as measured with single-cell recording

Stimulus-specific repetition-related reductions in firing rates have been found in physiological recordings of neurons in macaque inferior temporal (IT) cortex [14,15,17–19] (Figure 1). These repetition effects have been reported for awake behaving animals performing various visual tasks (e.g. match to sample [17], recognition memory [19,29]), as well as in anesthetized animals [30], and occur for both behaviorally relevant and irrelevant...
stimuli [2]. RS effects are stimulus specific in the sense that they do not appear to reflect a global reduction in the firing of a population of neurons to all subsequent stimuli [17]. Nonetheless, the precise definition of a ‘stimulus’ is important, because neural RS can exhibit invariance to some changes in stimulus dimensions (such as the size or position of an object [24]).

The degree of RS depends on several factors, although the precise effect of each factor can vary with brain region. RS persists despite many intervening stimuli, particularly in more anterior regions of IT cortex [2,31]. RS increases with more repetitions of the same stimulus, such that firing rates resemble an exponentially decreasing function of presentation number [1]. This reduction in firing rates occurs primarily for visually excited neurons, [14] and the greatest reduction tends to occur for neurons that were most active on the first presentation [1]. Importantly, several studies indicate that RS onsets rapidly, as fast as 70–80 ms in some perirhinal neurons [31] and with a mean population latency of around 150 ms in IT [15]. Indeed, Xiang and Brown [31] suggested that these effects are too fast for ‘top-down’ influences.

An important concern for the present discussion is the proportion of neurons showing RS. The best available estimates suggest that significant RS occurs for ~50–67% of visually responsive neurons in IT cortex [2,15]. There are two possible interpretations of these data. One is that they reflect two different neural populations: One showing RS, and the other showing no reduction in response [19]. Alternatively, there might be a single population, in which all neurons show some degree of RS (with the failure to observe RS in some neurons simply the result of limited power to detect a significant reduction). However, the exact relationship between RS and the degree of stimulus specificity has not been determined. This is a crucial issue that we will return to later (see section below: ‘Distinguishing the neural models’).

Repetition suppression as measured with fMRI

The basic phenomenon of RS measured with fMRI, also referred to as fMRI-adaptation, has been replicated many times in ventral temporal cortex (see review [32]) as well other areas such as medial temporal [6] and frontal cortex [33–35]. Many adaptation paradigms have been used to measure RS, including multiple repetitions of the same stimulus without intervening items [7] (Figure 2a: block design), or after a single presentation with either no [36] or many [37] intervening items (Figure 2a: rapid event-related). Thus, the properties of RS listed below might vary not only across different brain regions, but also as a function of the paradigm and task [38] (Box 2).

As with single-cell data, fMRI responses tend to decrease monotonically with the number of repetitions [9,16,39], often reaching a plateau after six to eight repetitions [8,39]. RS is maximal when there are no intervening stimuli between repeats [13,39], but has also been observed with tens of intervening stimuli [9] and even after multiple days between presentations [11,35,40]. Other factors that affect the level of RS are stimulus contrast [41], attention [42–44] and the duration of the initial stimulus presentation [45]. RS has also been reported for subliminal stimuli [10]. Some studies have argued that RS occurs across a wide expanse of object-responsive cortex, and not only in regions within that cortex that show maximal responses to particular stimuli [40,41]. However, the relationship between RS and stimulus selectivity has not been tested directly.

The first study to use fMRI-adaptation to probe the sensitivity of object representations to object transformations (size, position, viewpoint, and illumination) was conducted by Grill-Spector and colleagues [7] (Box 1). They found that the posterior section of lateral occipital cortex (LO), which is known to be object selective, showed sensitivity to many object transformations including size and position; however, more anterior object-selective regions were to some extent invariant to size and position, but sensitive to object rotation or illumination. A later
The different levels of rotation invariance found across studies could depend on the objects used, the degree of rotation and the type of paradigm used.

More recently, researchers have used fMRI-adaptation paradigms with monkeys [47–49]. This provides an important link between methodology (single-cell vs. fMRI) and species (monkeys vs. humans). For example, Sawamura et al. [49] found comparable size invariance, as inferred via fMRI-adaptation, for objects presented to both monkeys and humans.

Repetition effects as measured with EEG/MEG
Repetition effects have also been studied by measuring changes in the electrical (EEG) or magnetic (MEG) field, usually recorded above the scalp. These effects reflect changes in the amplitude and/or synchrony of local field potentials (LFPs) caused by transmembrane currents in large numbers of neurons. Most EEG studies examine event-related potentials (ERPs), which reflect changes in electrical potential during the few hundred milliseconds following stimulus onset, averaged across trials. The earliest object repetition effects are typically observed ~200 ms after stimulus onset [13,50–52]. Henson et al. [13], found effects of repeating the same view of an object as early as 160–190 ms when there were no intervening objects; with one or more intervening objects, repetition effects only emerged from ~200 ms. Schendan and Kutas [50] showed that repetition produced an ERP difference between 140 and 250 ms that was greater for objects repeated with the same view than with a different view, whereas later time windows (400–700 ms) showed repetition effects with greater generalization across views. ERP studies using long-lag repetition of faces, including studies that record directly from the cortical surface (in patients undergoing surgery [53]), tend to find that repetition effects emerge only after the first face-specific ERP component, the ‘N170’ [54,55]. Together, these studies suggest that when there are intervening stimuli between object repetitions, repetition effects on ERPs tend to occur soon after initial stimulus-specific components (>200 ms; Figure 3a), unless there is a change in view, in which case the effects emerge later (>400 ms).

Other EEG studies concentrate on changes in the power of electrical or magnetic oscillations that are induced by stimulus repetition (high-frequency oscillations are not observed in ERPs if they are not phase-locked across trials). Some studies report decreased high frequency (>40 Hz) power around 220–350 ms for repetition of familiar objects across lags of one or two intervening objects [52] (Figure 3b). Such changes in power in certain frequency bands have been shown to correlate with the blood oxygenation-level-dependent (BOLD) changes measured by fMRI [56].

Recent methods for distributed source localization of high-density EEG/MEG data provide estimates of spatiotemporal patterns of activity over the cortex and suggest that prefrontal activity might precede repetition effects in ventral temporal cortex. For example, Dale and colleagues [12] estimated that the onset for repetition effects (for words) recorded by MEG was ~250 ms in ventral cortex, which was after activity had reached frontal cortex.

The study by Vuilleumier and colleagues [37] used an event-related paradigm and reported similar results regarding size invariance, but found differential sensitivity to object rotation across hemispheres: the left hemisphere displayed invariance to object rotation, the right did not. A related study by James et al. [46] found higher invariance to object rotation in the ventral than the dorsal stream.
Box 2. Experimental assay of repetition suppression (RS) as measured by BOLD

Several procedural issues need to be kept in mind when comparing RS effects across different studies: the number of repetitions; time between repetitions; the paradigm used (prolonged repetition, i.e. adaptation, versus brief presentations); whether repeated stimuli are compared with initial presentations of the same stimuli or with initial presentations of different stimuli; whether the repetitions are task relevant or irrelevant [27]. RS is a relative measure, so it is important to remember that even the baseline condition (e.g. a series of different, non-repeated objects) might have some degree of RS owing to shared stimulus properties.

It is also important to distinguish studies that examine RS during intermixed initial and repeated presentations (event-related design) from those that compare the mean activity during a block of stimuli (repeat vs. non-repeat blocks). Block designs have the advantage of prolonged repetition that increases the magnitude of RS. However, additional factors, such as attention and expectation, might also contribute to differences in the BOLD signal. Thus it is important to control these factors across conditions. Subliminal priming paradigms overcome many of these problems [10], although subliminal effects tend to be smaller (and might be qualitatively different [50]).

Other designs compare the responses of two stimuli presented sequentially, in which the pair consists of either two presentations of the same or different stimuli [56]. This approach combines the advantage that short-lag repetition effects tend to be larger with the advantage of a reduced predictability of repetition (if the order of ‘same’ and ‘different’ pairs is randomized). However, it has a potential disadvantage that the RS effect might be a small proportion of the total BOLD response to both stimuli within a pair. Further, depending on the interval between stimuli, one might have to consider additional factors such as potential non-linearities in the relationship between neural activity and BOLD response at short interstimulus intervals.

(≈ 185 ms). A recent study that localized broadband (4–50 Hz) MEG power reported increased power for repeated vs. first presentations of objects from 120–230 ms in frontal cortex, followed by decreases from 350–440 ms in fusiform cortex [57]. These studies indicate multiple repetition effects across both space and time. They also suggest that some (long-lag) repetition effects in ventral temporal cortex might arise relatively late (contrary to the single-cell data), perhaps after interactions with frontal cortex.

Figure 3. Object-repetition effects measured with EEG. (a) ERPs from posterior left and right electrodes, in response to line drawings of familiar objects, repeated with 1–2 intervening objects. Data averaged across 11 participants. Black: first presentation of the stimulus; Red: repeated versions of the same stimulus. The difference between repeated and non-repeated curves is only reliable by the third time-window (230 ms onwards); the early deflections within the first two time-windows did not show significant repetition effects in this study. (b) Amplitude of EEG as a function of time and frequency for first (left plot) and repeated (right plot) presentations, using same paradigm as in (a). Power spectrograms were calculated on the basis of single trials, after subtracting the mean evoked response, and therefore reflect components that are not phase-locked to stimulus onset (see [52] for more details). Note the decrease in amplitude (color change towards green) from 220–350 ms (black box) for repeated relative to first presentations. This decrease occurs in the frequency band above 30 Hz (gamma band). (Data reproduced with permission from [52].)

Relating different types of data
Relating the data recorded from single cells by electrodes (firing rates, multi-unit activity: MUA, or LFP) to the data recorded by fMRI (BOLD) and by EEG/MEG (synchronous changes in LFPs over many neurons) is non-trivial. In general terms, increases in MUA correlate with increases in LFP, and increases in LFP correlate with increases in BOLD [58] and presumably EEG/MEG power. However, the special constraints of each technique give rise to situations where their measures might dissociate. For example, single-cell recording is biased towards large excitatory neurons and might not capture changes in smaller cells. Moreover, changes in LFPs from presynaptic activity might not always produce changes in postsynaptic firing (e.g. if excitatory and inhibitory presynaptic activity cancel out). Thus, in principle, presynaptic activity in neurons could increase, causing increases in BOLD and MEG/EEG power, even if the firing rates of neurons in that region did not. Furthermore, changes in the synchrony but not amplitude of LFPs could, in principle, be reflected in EEG/MEG signals, but not BOLD. Conversely, changes in LFPs in neurons that are not aligned might have detectable BOLD signals without detectable EEG/MEG signals. In addition, because the BOLD signal integrates over several seconds, transient effects or small shifts in the latency of neural activity might be difficult to detect with fMRI. Finally, issues concerning experimental design need to be taken into account when comparing across methodologies or even within a methodology (Box 2).

A related issue is that repetition effects in one region might change the balance of inputs this region sends to a second region, thus producing effects beyond the region that showed RS. For example, Tolias and colleagues [59] found that neurons in macaque visual area V4, which are not selective for motion direction, become selective for motion direction after adaptation to motion [47]. They
suggest that V4 neurons normally receive balanced inputs from motion-selective neurons in the middle temporal area, MT, but that after adaptation the input from MT becomes unbalanced, resulting in motion-selective responses in V4. This finding is relevant for the use of RS to infer the nature of representations in different brain regions (Box 1).

Models for repetition suppression

There are multiple potential neural causes of the RS measured with single-cell recording, fMRI or EEG/MEG. We now describe three models that have been sketched previously in the literature (Figure 4): (i) the Fatigue model, whereby the amplitude of firing of stimulus-responsive neurons decreases [2,16]; (ii) the Sharpening model, whereby fewer neurons respond [1,18,26]; and (iii) the Facilitation model, whereby the latency [60] and/or duration of neural activity is shortened [3,27]. An important consideration to keep in mind is how each of these models increases processing efficiency.

Fatigue model

According to this model, all neurons initially responsive to a stimulus show a proportionally equivalent reduction in their response to repeated presentations of the same stimulus. As a consequence the mean population firing rate declines but there are no changes in the pattern of relative responses across neurons, or in the temporal

---

**Box 3. Potential neural mechanisms underlying repetition suppression (RS)**

At least four mechanisms might play a role in RS (i.e. the reduction in spiking rates following repetition) at the neural level:

(1) **Firing-rate adaptation.** This reflects reduced excitability of neurons owing to an increase in potassium ion currents that hyperpolarize the membrane potential and increase its conductance [61]. This in turn reduces the impact of synaptic inputs and hence lowers the probability of neural spiking. This effect tends to be short lived, decaying over hundreds of milliseconds, although possibly lasting tens of seconds (when potassium currents are activated by sodium ions [90]).

(2) **Synaptic depression.** This is a temporary reduction in synaptic efficacy that is believed to reflect reductions in presynaptic neurotransmitter release [91]. The effect depends on pre- (rather than post-) synaptic activity and can last tens of seconds for excitatory synapses (it tends to be shorter-lived for inhibitory synapses).

(3) **Long-term depression (LTD).** This longer-lasting mechanism reflects plasticity changes owing to correlated pre- and postsynaptic activity and normally entails multiple stages, including decreases in calcium concentration, perhaps even gene expression and ultimately protein synthesis. LTD reflects a reduction in synaptic efficacy [92], normally follows low-frequency stimulation and can last at least hours.

(4) **Long-term potentiation (LTP).** Like LTD, this long-lasting mechanism entails multiple stages and might come in several forms (depending on the precise receptor type). LTP reflects an increase in synaptic efficacy. Both LTP and LTD have been best studied for excitatory (glutamatergic) synapses, for which they tend to increase or decrease, respectively, the probability of spiking following presynaptic input. Although controversial, similar mechanisms might also operate on inhibitory synapses (from interneurons [93]), in turn causing either decreases or increases in spiking probability respectively.
Sharpening model predicts that repetition will narrow the tuning bandwidth, that is, sharpen the response around the preferred stimulus. 

This mechanism does not explain the specificity of RS, that is, why the neuron's response is reduced to some stimuli, yet resumes high firing rates to other stimuli. An alternative mechanism is reduced synaptic efficacy of specific synapses from connected neurons (synaptic depression; Box 3). In this manner, only specific patterns of presynaptic input to the neuron (which are stimulus dependent) reduce its firing rate. This type of mechanism has been implicated in early visual cortex and usually occurs with prolonged repetitive stimulation.

One prediction from this model is that the amount of RS will be greater in neurons that respond optimally to a stimulus than in other neurons (Figure 5a). As a result, the sensitivity of the system to stimuli that are different from the repeating stimulus is increased, thereby providing a mechanism for 'novelty detection'. Reducing the firing rate might also help prevent saturation of the neural response function by increasing its dynamic range. This type of mechanism has been implicated in early visual cortex and usually occurs with prolonged repetitive stimulation.

Another advantage hypothesized for such a mechanism is that it reduces redundancies in the neural code, which increases the efficiency of information encoding [62]. However, this mechanism does not explain the specificity of RS, that is, why the neuron's response is reduced to some stimuli, yet resumes high firing rates to other stimuli. An alternative mechanism is reduced synaptic efficacy of specific synapses from connected neurons (synaptic depression; Box 3). In this manner, only specific patterns of presynaptic input to the neuron (which are stimulus dependent) reduce its firing rate. This type of mechanism has been implicated in early visual cortex and usually occurs with prolonged repetitive stimulation.

One prediction from this model is that the amount of RS will be greater in neurons that respond optimally to a stimulus than in other neurons (Figure 5a). As a result, the sensitivity of the system to stimuli that are different from the repeating stimulus is increased, thereby providing a mechanism for 'novelty detection'. Reducing the firing rate might also help prevent saturation of the neural response function by increasing its dynamic range. Another advantage hypothesized for such a mechanism is that it reduces redundancies in the neural code, which increases the efficiency of information encoding [62].

However, it is not immediately clear how reduced firing rates can account for increased speed and accuracy of processing repeated stimuli (in addition to increased sensitivity to novel stimuli), as arises in priming. One explanation is provided in a computational model by Gotts (S. Gotts, PhD thesis, Carnegie Mellon University, USA, 2003), in which a reduction in the mean and variance of firing rates allows for greater synchrony of neural responses. Greater synchrony of presynaptic input is believed to be more effective in triggering a postsynaptic response [63], which would imply more rapid transmission of information throughout the network, resulting in faster responses (priming). A key prediction of this model is that synchrony can increase while stimulus-specific firing rates decrease. Although there is evidence in support of this possibility [64], others have argued for the opposite effect [65]. An increase in synchrony might also be difficult to reconcile with observations of reduced oscillatory power following stimulus repetition [52], although it is possible that decreased amplitude of local field potentials outweighs the increased synchrony of these potentials.

**Sharpening model**

Desimone and colleagues [18] and Wiggs and Martin [26] have suggested that repetition results in sparser representation of stimuli. According to this model, some – but not all – neurons that initially responded to a stimulus will show RS to subsequent presentation of that stimulus. Importantly, it is the neurons that code features irrelevant to identification of a stimulus that exhibit RS (Figures 4,5b). Thus, repetition-related changes are viewed primarily as a learning process, in which representations (e.g. tuning curves, Figure 5b) are 'sharpened' and, as a consequence, the distributed representation becomes sparser, resulting in fewer responsive neurons in total. An important difference between the Sharpening and Fatigue models is that, for Sharpening, many of the neurons that are optimally tuned to the repeating stimulus should show little or no response reductions, rather than exhibit the greatest response reduction, as in the Fatigue model.

Sparser representations clearly have adaptive value in terms of a reduced metabolic cost. Also, because the representation becomes sharper (tuning curves become narrower, Figure 5b), the neurons become more sensitive to change. Sparser representations might also allow for more efficient or faster processing, although this depends on the manner in which their information is read-out by downstream neurons. Because the Sharpening model suggests a changed and improved representation for repeated stimuli, this model has been widely used to explain priming [26,27,45]. However, a recent study suggests that some RS in object-selective cortex might reflect response learning and implies that object representations do not necessarily reorganize as a consequence of repetition [66].

The mechanism underlying the formation of sparser representations is unknown but could reflect inhibition from lateral connections between neurons within...
a population. For example, Norman and O’Reilly [67] used a competitive Hebbian learning rule to simulate the sharpening of representations with repetition (within medial temporal cortex). Initially, many neurons respond weakly to a distributed input pattern representing the stimulus. Through competitive interactions, the neurons with the strongest initial response become ‘stronger’, and inhibit the ‘weaker’ neurons. Thus some neurons show increased firing rates following repetition, whereas others show decreased firing rates. By assuming that the number of ‘strong’ units is less than the number of ‘weak’ units, the population response decreases with repetition because there are more neurons showing reduced activity than showing increased activity. If information from only those neurons with high firing rates is ‘read out’ by downstream neurons, their increased firing rate following repetition (despite the global decrease) could explain the faster processing of repeated stimuli.

**Facilitation model**

At its simplest, this model predicts that repetition causes faster processing of stimuli, that is, shorter latencies or shorter durations of neural firing. One example is the ‘accumulation’ model of James et al. [60], in which stimulus information is accrued faster following repetition. Given that the hemodynamic signal measured by fMRI integrates over several seconds of neural activity, a shorter duration of activity results in a decreased amplitude of the fMRI signal. A shorter duration of neural activity is also consistent with earlier peaking of the fMRI response [68], and might explain why decreases in firing rate can appear to arise after the initial visual response [15]: the neurons initially fire robustly to both first and repeated presentations, but this firing stops sooner for repeated presentations.

An extension of the Facilitation model assumes that the cause of this faster processing is synaptic potentiation between neurons following an initial stimulus presentation, and that this potentiation can occur at many levels in the processing stream. As a consequence, information flows through the stream more rapidly, and hence identification of a repeated stimulus occurs faster. In terms of attractor neural network models, synaptic potentiation might be viewed as deepening the basin of attraction, resulting in a shorter time for the network to settle on a stable pattern corresponding to identification of the stimulus. An example of such a dynamical, network model is sketched by Friston [69]. The key idea behind this model is that the firing rate of the long-range excitatory (output) neurons in a population codes ‘prediction error’ [70], which is the difference between bottom-up input (‘evidence’) and top-down input (‘prediction’). The dynamics of the network are such that prediction error decreases over time after stimulus onset, and synaptic changes serve to accelerate this decrease when the stimulus is repeated (i.e. repetition improves prediction).

This emphasis on recurrent activity between many levels of the processing stream is consistent with the spatiotemporal pattern of repetition effects emerging from MEG/EEG data. Moreover, if interregional interactions require an initial volley of activity through the network [71], this model could further explain the relatively late onset of long-lag repetition effects recorded with EEG/MEG. However, such a model would not explain why much earlier repetition effects have been observed in some neurons (e.g. 75–100 ms, which is thought to be too early for feedback [31]); further, this model does not necessarily predict decreases in the peak firing rate of individual neurons.

Each of the above models would clearly benefit from further elaboration, including instantiation as detailed computational models. It is possible that different models might apply in different brain regions and under different experimental conditions (e.g. different paradigms/tasks). We next consider some experiments that might distinguish between the models.

**Distinguishing the neural models**

There are three main directions in which these models can be distinguished: (i) examining the relationship between RS and stimulus selectivity; (ii) examining the effect of repetition on the tuning of cortical responses along a stimulus dimension; and (iii) examining the temporal window in which processing occurs for new and repeated stimuli.

**Examining the relationship between RS and stimulus selectivity**

The models differ in their predictions on whether RS is strongest for the preferred stimulus, or for non-preferred stimuli. The Sharpening model predicts that neurons showing little or no RS to a repeated stimulus are highly selective for that stimulus. By contrast, both the Fatigue and Facilitation models predict that RS is proportional to the initial response. Thus, neurons that respond optimally for a stimulus should show the largest suppression. These hypotheses can be tested with single-cell recording.

**Examining the effect of repetition on neural tuning**

Another way to distinguish the models would be to find a single dimension (e.g. motion, orientation), along which stimuli differ, and examine the effect of repetition on the tuning curves of different neurons along that dimension. The models differ in their prediction of how repetition will change the shape of neuronal tuning. According to the Fatigue model, repetition reduces the response in proportion to the initial response, but the tuning width does not change (Figure 5a). Most probably, the reduction will be maximal for tuning curves centered on the location of the repeating stimulus along the stimulus dimension, and lesser for tuning curves centered further away. This is consistent with physiological data in V1 for orientation [72], spatial frequency [73] and motion direction [74]. By contrast, according to the Sharpening model, repetition ‘sharpenst’ tuning curves (Figure 5b). This is consistent with studies of learning-related changes in IT cortex and V4 [75–77]. Finally, the Facilitation model does not suggest any particular effect on tuning curves (Figure 5c). Indeed, even a widening of the curves might be possible if repetition enlarged the attractor basin in an attractor network model.
An elegant study by Kohn and Movshon [74] demonstrates how measuring tuning to a stimulus dimension before and after repetition can be used to distinguish between models. Kohn and Movshon examined the effect of repetition (using prolonged adaptation) to the preferred, non-preferred and null motion direction on tuning curves of MT neurons. Their data reveal that the tuning of MT neurons to motion direction narrows (sharpens) around the repeating stimulus when adapting a cell with its preferred motion direction and, further, that tuning curves shift towards the repeating stimulus when adapting the neurons with a non-preferred stimulus. However, sharpening of tuning curves did not occur in V1; rather V1 data were consistent with the Fatigue model. These findings provide evidence that stimulus repetition can have different neural consequences depending on brain region. Thus, different models of RS could apply for different cortical sites.

Examining the temporal window of processing for new and repeated stimuli

The models can also be distinguished in the temporal domain. In particular, the Facilitation model suggests that the latency and/or duration of the response to repeated items will be shorter than to first presentations. The Fatigue and Sharpening models do not suggest a difference in the temporal processing window for repeated stimuli. The latency and duration of processing might be examined via single-cell recordings and/or EEG/MEG techniques.

Another prediction from one version of the Facilitation model is that facilitation will be more pronounced after top-down information is processed. As a result, the largest difference between repeated and non-repeated stimuli should occur after initial sensory processing (e.g. after the first 200 ms). This is consistent with the observation that long-lag repetition effects typically emerge after initial stimulus-specific ERP/ERF components. Unlike the Facilitation model, the Fatigue model suggests that RS occurs during the entire processing window. This predicts that the earliest selective responses are modulated by repetition. For the Sharpening model, the onset of repetition effects might depend on whether sharpening is implemented in terms of changes in feedforward (input) synapses, or via changes from lateral connections within a region. In the latter case, RS might be delayed until selective neurons have inhibited non-selective neurons, which would be consistent with RS occurring subsequent to the initial stimulus-specific response.

We would like to emphasize that any empirical data relevant to the above predictions are likely to depend on other factors, such as the lag between repetitions. One of the central outstanding questions is whether different models apply at different time scales. One possibility is that the mechanisms related to the Fatigue model operate during immediate repetitions of a stimulus within a few hundred milliseconds, and reflect transient stimulus-specific effects that onset rapidly. By contrast, the effects of repetition across many intervening stimuli might be more consistent with the Sharpening or Facilitation models, and reflect long-term learning that leads to changes in the spatial pattern of stimulus-selective responses and/or dynamics of those responses.

Conclusions

Adaptation paradigms have become increasingly popular for examining the processing characteristics of different cortical regions of the brain. Clearly, progress has been made using RS to infer the nature of representations in different cortical regions, or as a marker for increased processing efficiency, without a complete understanding of its neural basis. Nonetheless, it is our belief that the specific neural mechanisms matter, because interpretation and design of experiments depend on the nature of the underlying neural model. For example, models differ as to whether RS reflects quantitative or qualitative changes in representations. One important possibility we raised is that of multiple models – which vary in their relevance across space, time and task – that might parallel the multiplicity of potential neural/synaptic mechanisms (Box 3). Although it is possible that a single model could apply under all conditions, we think it is likely that different models are needed to explain RS in different conditions. Some effects might be short-lived, others might be permanent. Further, the magnitude of RS might vary as a function of task and stimuli. Finally, it is not yet known whether the same or different mechanisms operate in different brain regions.

Clearly, many questions remain regarding the neural basis of RS (Box 4). Progress will be aided primarily by a better understanding of the effects of stimulus repetition on single neurons. This will include concurrent recording from multiple electrodes to measure firing rates and their correlations between neurons. Progress can also be made using more global measures in humans, such as fMRI and EEG/MEG, provided important differences between these measurements are kept in mind. Improvements in the spatial resolution of fMRI [78] are likely to be important for distinguishing between models. The use of similar stimuli and paradigms in physiological experiments and fMRI/EEG/MEG experiments will be key for comparisons across different types of data. Most informative will be concurrent recording of repetition effects using all these methods. These future experiments will yield important empirical data that will validate (or refute) the theoretical predictions laid out in this review.

Box 4. Questions for future research

- What is the relationship between repetition suppression (RS) and stimulus selectivity?
- How does repetition affect the tuning curves of neurons?
- How do the Fatigue and Sharpening models account for improved behavioral performance (priming)?
- In what time scale does facilitation occur?
- Does facilitation require interactions between different brain regions, in particular feedback?
- Does RS induced by prolonged adaptation and RS induced by one (or several) presentations with intervening stimuli involve similar or distinct neural mechanisms?
- How does RS in lower level regions (e.g. V1) differ from RS in high-level regions (e.g. MT or LO)?
Acknowledgements
The authors have contributed equally to this manuscript. We would like to thank Steve Gotts for many helpful conversations, Rory Sayres and Sven Heinrich for contributing to Figure 2, and Thomas Gruber for contributing Figure 3. We would also like to thank Dave Andresen, Sharon Gilai-Dotan, Sven Heinrich, David Leopold, David McMahon, Rafael Malach, Mortimer Mishkin, Carl Olson, Leslie Ungerleider, Anthony Wagner, and Nathaniel Wittoff for their insightful comments on the manuscript. Finally, we would like to thank the anonymous reviewers for their comments on the manuscript.

References

Endeavour

Coming soon in the quarterly magazine for the history and philosophy of science:

**The accidental conservationist** by M.A. Andrei
**Powering the porter brewery** by J. Sumner
**Female scientists in films** by B.A. Jones

Locate Endeavour on ScienceDirect (http://www.sciencedirect.com)