The Operation of Selective Attention at Multiple Stages of Processing: Evidence from Human and Monkey Electrophysiology

STEVEN J. LUCK AND STEVEN A. HILLYARD

ABSTRACT Because of its finite computational resources, the human brain must process information selectively in a variety of domains. For example, we may limit processing to a subset of the many possible objects that could be perceived, a subset of the many possible memories that could be accessed, or a subset of the many possible actions that could be performed. Although the fundamental need for selective processing is present at each of these stages, there are important differences in the computations performed by different cognitive systems; hence it is likely that substantially different attentional mechanisms are responsible for selective processing at each stage. It has been difficult, however, to isolate the attentional mechanisms that operate in different cognitive systems, primarily because behavioral outputs typically reflect the combined effects of several stages of processing. In this chapter, we describe the application of electrophysiological techniques to the isolation of attentional operations at different stages of processing, including perception, working memory, and response selection.

During almost every waking moment, the human brain is confronted with a vast amount of information—thoughts, memories, and emotions—and innumerable sensory inputs from every modality. To avoid being overwhelmed by this cacophony, the brain relies on mechanisms of selective attention that limit processing to a subset of the available information. The resulting selectivity is easily observed in everyday human behavior: We generally focus our thoughts and conversations on a small number of topics at any given moment, and we typically notice and remember only a fraction of the objects and events we encounter. Moreover, because we have but two hands, two feet, and one mouth, our actions are typically directed toward a small number of tasks at any one time.

Although the effects of selective attention on behavior are ubiquitous, the internal mechanisms by which information is selected are just beginning to be understood. Indeed, the pervasiveness of selectivity is partly responsible for the difficulty of understanding how attention operates, because a given behavioral sign of selective attention could potentially reflect internal selectivity within any of the various cognitive subsystems that underlie a given behavior. Over the past 25 years, the techniques of cognitive neuroscience have played an increasingly important role in helping to analyze the level of processing at which information is selected and to understand the neural substrates of the selection process.

In this chapter, we focus on two fundamental and interrelated issues in the study of attention. The first issue is the classic “locus-of-selection” question, which asks whether attention operates at an early, perceptual stage or at a late, postperceptual stage. Previously, we have stressed the operation of attention at early stages (Hillyard et al., 1995), but here we discuss conditions under which attention operates at intermediate and late stages as well as at early stages. The second issue to be addressed in this chapter is the question of exactly why selective attention is needed at certain stages of processing. For many years, psychologists were satisfied with the proposal that cognitive processes were simply limited in capacity and that attention was used to allocate some sort of processing resource. But it is now possible to specify more exactly how selective attention may solve specific computational problems that arise in cognitive processing, such as the so-called “binding problem,” which occurs when independently coded attributes of an object must be combined (Luck, Girelli, et al., 1997; Treisman, 1996).
Electrophysiological techniques

Several neuroscience techniques can be used to study attention, but in this chapter we focus primarily on event-related potentials (ERPs) in humans and single-unit recordings in monkeys. ERPs are voltage deflections that can be recorded on the scalp at the time of a stimulus, a response, or some other measurable event. These voltage deflections typically arise from summated postsynaptic potentials in the cerebral cortex that are passively conducted through the brain and skull to the scalp, where they can be recorded noninvasively from normal volunteers (reviewed in Hillyard and Picton, 1987). As illustrated in figure 47.1, the ERP waveform following a stimulus consists of several positive (P) and negative (N) deflections called “waves” or “components.” The sequence of components following a stimulus reflects the sequence of neural processes triggered by the stimulus, beginning with early sensory processes and proceeding through decision- and response-related processes. The amplitude and latency of the successive peaks can be used to measure the time course of cognitive processing, and the distribution of voltage over the scalp can be used to estimate the neuroanatomical loci of these processes. ERPs have two main advantages over traditional behavioral measures in the context of attention research. First, they provide a continuous measure of processing between a stimulus and a response, making it possible to determine the stage or stages of processing that are influenced by attention. Second, ERPs provide a means of covertly measuring the processing of a stimulus—without requiring an overt response—which is very useful for assessing the processing of unattended stimuli. However, the spatial resolution of the ERP technique is limited, and so we also discuss data obtained from other human neuroimaging techniques and from recordings of single-unit activity in monkeys.

Multiple stages of selection

Across a wide variety of conditions, it can be shown that people are faster and more accurate at detecting attended stimuli than unattended stimuli. However, the origins of such effects have been debated since the 1950s, with some researchers arguing that attended stimuli are better perceived than unattended stimuli (the “early-selection” position) and others arguing that perceptual processing is unaffected by attention and that selective processing occurs only at postperceptual stages, such as working memory and response selection (the “late-selection” position). It is beyond debate that attention does sometimes operate at a late stage—after all, there are clearly occasions when an individual fully perceives a stimulus but chooses not to make an overt response to it. The burden of proof therefore rests on the shoulders of those who argue that attention operates at an early stage. As we discuss in the next section, neurophysiological studies have now conclusively demonstrated that attention operates at early stages as well, at least under some conditions. Thus, attention may oper-

![Diagram](image)

**Figure 47.1** Stimuli and results from a typical ERP attention experiment (Clark and Hillyard, 1996). In experiments such as this, the subject fixates a central point and attends to either the left or right visual field. Stimuli are flashed to the left and right fields in a rapid, randomized sequence, and the subject responds to occasional targets embedded among frequently occurring nontargets in the attended field. (A) Grand average ERP waveforms from four different scalp sites, showing the C1, P1, and N1 components. The P1 and N1 components were larger for attended-location stimuli than for unattended-location stimuli, but the C1 component was unaffected by attention. The scalp map shows the distribution of voltage from 60–85 ms. In this time period, the C1 wave appears as a negative potential (indicated by light gray) at the midline, and the P1 wave appears as a positive potential (indicated by dark gray) at lateral occipital sites. (B) Estimates of the C1 and P1 dipole sources, projected onto coronal brain sections derived from the atlas of Talairach and Tournoux (1988). These best-fit dipoles were calculated from the scalp voltage distributions using the BESA algorithm (Scherg, 1990). (Reprinted with permission from Hillyard, Teder-Sälejärvi, and Münte, 1998.)

688 ATTENTION
ate at either early or late stages of processing (or both), and the task for attention researchers has now shifted to determining what conditions lead to attentional selection at a given stage of processing.

We propose that a simple principle determines the stage at which attention operates: The presence or absence of attentional selection at a given stage of processing depends on the presence or absence of interference at that stage, which in turn depends on the stimuli and task. For example, consider a task in which red and green letters are presented individually at fixation at a rate of one letter per second, and the subjects are required to remember all the red letters, reporting them at the end of a 15-second trial block. The visual system could easily identify both the red letters and the green letters at this slow stimulus presentation rate; therefore, no selective processing of the red letters would be expected at the stage of perception. However, the limited capacity of working memory would make it difficult for most subjects to retain both the red and the green letters, so only the red letters would be selected for storage in working memory. In contrast, a very different pattern of selectivity would be expected if 14 green letters and 1 red letter were presented simultaneously in a densely packed array and the subject were required to report the identity of the red letter as quickly as possible. In this case, the memory load would be relatively modest, but the green letters might interfere with the perception of the red letter, thus necessitating a perceptual-level filtering of the green letters (i.e., early selection). Perceptual-level interference may also arise with individually presented stimuli under certain conditions, such as when the stimuli are presented at such low intensities that spontaneous neural noise becomes a significant source of interference (Hawkins et al., 1990).

By this account, the locus of selection is directly determined by the stage of processing at which interference is present for a given stimulus-task combination (for a related account, see Lavie, 1995). Although this is a useful conceptualization of the locus-of-selection issue, it is difficult to test unless there is an independent means of determining the conditions under which interference will be present at a given stage. In the following sections, we describe conditions that would be expected to lead to interference at three coarsely defined stages of processing—namely, perception, working memory, and response selection.

**Evidence for selection at the perceptual level**

**The Onset Time of Selective Processing** The application of the ERP technique to the locus-of-selection issue relies on straightforward logic: If attention operates at an early stage, then the early sensory-evoked ERP waves should be enhanced for attended stimuli relative to unattended stimuli, whereas if attention operates at a late stage, then only the later ERP waves should be affected. Several studies have used this logic in combination with variants of the experimental paradigm illustrated in figure 47.1A. In this paradigm, subjects fixate a central point while stimuli are flashed at predefined locations in the left visual field (LVF) and the right visual field (RVF). Subjects are instructed to attend to the LVF in some trial blocks and to the RVF in other trial blocks, and they are required to respond when they detect an infrequent target stimulus that occurs occasionally among the nontarget stimuli at the attended location. The stimuli are presented rapidly, making it difficult for the subjects to detect the targets without the use of focused attention. Under these conditions, numerous studies have shown that the initial P1 and N1 waves are larger in amplitude for stimuli presented at the attended location than for stimuli presented at the unattended location (reviewed in Hillyard, Vogel, and Luck, in press).

There are several reasons to believe that the modulation of P1 amplitude by spatial attention reflects a modulation of perceptual rather than postperceptual processing. First, the P1 component is an obligatory sensory response from the occipital cortex (see below) that is sensitive to stimulus factors such as contrast and stimulus position and is insensitive to cognitive factors (with the exception of spatial attention). Attentional modulations of this component are therefore most readily attributable to changes in sensory processing (Johannes et al., 1995; Wijers et al., 1997). Second, this effect typically begins 70–90 ms after stimulus onset, at which time visual information is only beginning to reach the object recognition areas of the inferotemporal cortex. Third, effects of this nature have been observed only in studies of spatial attention and have not been observed when attended and unattended stimuli are presented at the same location and differ in other stimulus dimensions such as color (Hillyard and Anllo-Vento, 1998). If attention operates after stimulus identification is complete, there would be no reason to expect that the dimension of spatial location would be treated any differently from other stimulus dimensions (e.g., see Bundesen, 1990), whereas a special mechanism for selection by location is entirely consistent with the topographically organized representations used in the early and intermediate stages of visual processing. Finally, equivalent P1 attention effects are observed regardless of the task-relevance of a stimulus. In particular, P1 amplitude modulations are observed both for targets and nontargets, and even for task-irrelevant “probe” stimuli that are occasionally flashed at the
attended and unattended locations but do not resemble the task-relevant stimuli (Heinze et al., 1990; Luck, Fan, and Hillyard, 1993). Presumably, any attention effects that operate after a stimulus has been identified would be limited to task-relevant stimuli, because there would be little point in attending to a stimulus already identified as task-irrelevant. In contrast, any attention effects that occur before identification is complete must necessarily be insensitive to the identity of the stimulus. It should also be noted that results of this nature have been observed in the visual search and spatial cuing paradigms, which are frequently used to assess the behavioral effects of selective attention (Luck, Fan, and Hillyard, 1993; Luck and Hillyard, 1995; Luck et al., 1994; Mangun and Hillyard, 1991).

**NEURAL SITES OF SELECTIVE ATTENTION** Although the P1 attention effect almost certainly represents a modulation of perceptual processing, it is not yet clear exactly which specific substrates of perception are being modulated or which specific areas of visual cortex are subject to attentional control. To provide more specific information about the locus of selection within perception, recent studies have attempted to determine the neuroanatomical sources of the ERP attention effects and have recorded single-unit activity in specific areas of visual cortex in monkeys.

Recordings of ERPs from midline parieto-occipital scalp sites typically include a component called C1 that has an onset latency of 50–60 ms, which is even earlier than that of the P1 (see figure 47.1A). Two sources of evidence indicate that the C1 wave is generated in primary visual cortex (V1). First, as shown in figure 47.1B, dipole modeling procedures have indicated that the C1 scalp distribution is consistent with a generator source in area V1. Second, the scalp distribution and polarity of the C1 wave change as a function of stimulus position in the visual field, and these changes correspond closely to the known retinotopic projections of the upper versus lower visual fields upon the striate cortex within the calcarine fissure (Clark, Fan, and Hillyard, 1995; Mangun, Hillyard, and Luck, 1993). Several studies have found that the C1 wave is unaffected by spatial attention under conditions that lead to significant modulations of P1 amplitude (Clark and Hillyard, 1996; Gomez Gonzales et al., 1994; Mangun, Hillyard, and Luck, 1993), which indicates that the initial sensory response in V1 is not affected by attention [although attention may influence V1 at a later point in processing (see Aine, Supek, and George, 1995)]. In contrast, the neural generators of the P1 wave can be accurately modeled by dipoles in extrastriate cortical areas (figure 47.1B), and the effects of attention on the P1 wave suggest that spatial attention first influences visual processing in extrastriate areas of visual cortex.

To assess the neural locus of the P1 attention effect more precisely, Heinze and colleagues (1994) conducted a study in which subjects performed a variant of the attention task shown in figure 47.1 during both an ERP recording session and a positron emission tomography (PET) session. The P1 attention effect was found to be associated with an increase in blood flow in the ventral occipital lobe in the posterior fusiform gyrus, and dipole modeling indicated that the scalp distribution of the P1 attention effect was consistent with a generator source in this region. Mangun and colleagues (1997) subsequently reported that both the P1 attention effect and the increase in fusiform blood flow are influenced similarly by task manipulations, thereby reinforcing the hypothesis that the two measures reflect the same underlying neural activity. A second area in the medial occipital gyrus also showed an effect of attention on blood flow. For stimuli presented in the lower visual field, however, Woldorff and colleagues (1997) observed increased blood flow and an associated P1 dipole in a dorsal extrastriate area of the occipital lobe, with a secondary focus in the fusiform gyrus. Although the specific areas affected by attention vary among the experiments, these combined PET and ERP data strongly indicate that early perceptual selection takes place in extrastriate visual cortex.

Converging evidence concerning the localization of attentional modulation was provided by a recent study in which single-unit activity was recorded from macaque monkeys while the animals performed a task similar to that shown in figure 47.1 (Luck, Chelazzi, et al., 1997). As shown in figure 47.2A, many neurons in area V4 exhibited larger responses to attended-location stimuli than to unattended-location stimuli. Figure 47.2B plots the difference in the response to the attended- and unattended-location stimuli, showing that the effect of attention began at 60 ms poststimulus, which is the same time as the beginning of the stimulus-evoked neural response. Thus, attention appears to operate as a preset filter that modulates the initial afferent volley of activity through area V4. In addition, as in the case of the P1 attention effect, this modulation of single-unit activity was identical for target and nontarget stimuli (see figure 47.2C). In contrast, no consistent effects of attention were observed in a parallel set of recordings from area V1, which is consistent with the lack of any attention effect for the C1 wave in ERP recordings. Together, the ERP, PET, and single-unit results provide compelling evidence that selective attention can modulate perceptual processing in extrastriate visual cortex, at least under cer-
Figure 472. Single-unit responses recorded from area V4 in the study of Luck and co-workers (1997a). (A) Average post-stimulus histograms from 37 neurons that showed a significant difference in the response to an attended-location stimulus compared to an unattended-location stimulus. These responses were elicited by nontarget stimuli. (B) Difference between the responses to attended- and unattended-location stimuli. Note that this difference deviates from zero at approximately 60 ms, which is the same time as the onset of the sensory response. (C) Responses to target stimuli presented at the attended and unattended locations. (Adapted with permission from Luck, Chelazzi, et al., 1997)

Conditions leading to early selection

As discussed above, we propose that attention will operate at a given stage of processing only when there is interference at that stage. Lavie and her colleagues have similarly argued that attention operates at an early stage only under conditions of high perceptual load (Lavie, 1995; Lavie and Tsai, 1994). Consistent with these proposals, early attentional modulations of neural activity have been observed only when a significant perceptual load was imposed by requiring speeded responses, by using difficult-to-detect targets, by surrounding the target with competing nontarget stimuli, or by presenting the stimuli at a very rapid rate (e.g., see Luck, Fan, and Hillyard, 1993; Luck et al., 1994; Mangun and Hillyard, 1991). As detailed below, these are conditions that are likely to create perceptual-level interference and are thus likely to lead to perceptual-level attention effects. In particular, there are at least two different types of interference arising at the stage of perception that may be alleviated by selective attention. The first of these is interference among multiple objects, which can lead to binding errors (i.e., errors caused by incorrectly combining features that belong to different objects), and the second is interference from spontaneous neural activity.

The Binding Problem

The problem of correctly binding together the features of an object can arise at many stages of the nervous system, and one variant of this problem is illustrated in figure 473. This figure shows the responses of a set of four idealized V4 neurons that are selective for different features but have overlapping receptive fields. When a red horizontal bar is presented (figure 473A), the red-selective and horizontal-selective neurons are activated and the stimulus is coded correctly. Similarly, when a green vertical bar is presented (figure 473B), the green-selective and vertical-selective neurons are activated and the stimulus is again coded correctly. However, when the red horizontal and green vertical bars are presented simultaneously (figure 473C), all four neurons are activated and the neural coding becomes ambiguous; in particular, it is not possible to determine from these neurons whether the display contains a red horizontal bar and a green vertical bar or a green horizontal bar and a red vertical bar. In the terms used by Treisman and Schmidt (1982), the presence of multiple stimuli within the receptive fields of these cells may lead to “illusory conjunctions.”

Luck, Girelli, and colleagues (1997) recently proposed an “ambiguity resolution theory” of attention, which postulates that ambiguity in neural coding can be resolved by an attentional mechanism that limits processing to a single object at any given moment. If the inputs from the other objects are momentarily suppressed, then the neurons that remain active must necessarily correspond to the attended object (figure 473D). The theory also proposes that such filtering will be necessary only under conditions that lead to ambiguous neural coding. For example, when an observer is required simply to de-
tect the presence of red in a stimulus array, there is no ambiguity because the response of the red-selective neurons unambiguously signals the presence or absence of the color red, independently of the other objects in the display [as long as the colors of the objects are easily discriminable (see Treisman and Gormican, 1988)]. In addition, a neuron's responses are ambiguous only when multiple stimuli are presented inside its receptive field; when only a single object falls inside the receptive field, the neuron's response unambiguously codes the features of that object (figure 473E).

Evidence supporting this view of attention has been obtained in several single-unit recording studies (Chelazzi et al., 1993; Luck, Chelazzi, et al., 1997; Moran and Desimone, 1985; Treue and Maunsell, 1996), one of which is summarized in figure 474. Panel A of this figure shows the responses obtained from an individual neuron in area V4. When the monkey performed a simple fixation task, this neuron gave a large response to a blue vertical bar presented alone and a small response to a green horizontal bar presented alone. When the two stimuli were presented together and the monkey was instructed to attend to one of the two stimulus locations, the response of the cell was modulated by attention such that the response elicited by both stimuli presented simultaneously was similar to the response to the attended stimulus presented alone (figure 474B). In other words, when the monkey attended to the blue vertical bar, the response to the pair of bars was large, just like the response to the blue vertical bar presented alone; when the monkey attended to the green horizontal bar, the response to the pair of bars was small, as was the response to the green horizontal bar presented alone. Thus, attention appeared to filter out the contribution of the unattended stimulus.

Figure 474C shows the effects of attention averaged over a population of 29 neurons in area V4. Again, these neurons gave a larger response when attention was directed to an effective stimulus (i.e., a stimulus that elicited a large response regardless of attention) than to an ineffective stimulus (i.e., a stimulus that elicited a small response regardless of attention). As shown in figure 474D, however, this effect was eliminated when only one of the two stimuli was inside the receptive field and the other stimulus was outside the receptive field.

The effects of attention have been assessed in areas V1, V2, V4, and MT of the occipital lobe and also in inferior temporal cortex (Chelazzi et al., 1993; Luck, Chelazzi, et al., 1997; Moran and Desimone, 1985; Treue and Maunsell, 1996). In all of these areas, the results have suggested that strong attention-related modulations of sensory activity occur primarily when multiple stimuli are presented inside the classical excitatory receptive field (or so close to the receptive field that they
would be expected to indirectly influence the response of the cell. Receptive fields are so small in area V1 that it is virtually impossible for an animal to attend to one stimulus inside the receptive field and ignore another that is also inside the receptive field, and attention effects have only rarely been observed in this area (Mottet, 1993). In area V2, receptive fields are only slightly larger, but attention effects have been observed in the subset of V2 cells with large enough receptive fields to contain both an attended and an unattended stimulus (Luck, Chelazzi, et al., 1997). In V4, most receptive fields are large enough to contain multiple stimuli, and two studies have found that the effects of attention are substantially larger when multiple stimuli are presented inside the receptive field (Luck, Chelazzi, et al., 1997; Moran and Desimone, 1985). A similar result has been reported for area MT (Treue and Maunsell, 1996). In inferotemporal cortex, the receptive fields are so large that it is difficult to place a stimulus outside the receptive field, and the majority of cells in this area exhibit large attention effects (Moran and Desimone, 1985). As will be discussed later, attention effects have sometimes been observed in extrastriate cortex in the absence of multiple stimuli inside the receptive field (Connor et al., 1996; Mottet, 1993; Spitzer, Desimone, and Moran, 1988; Treue and Maunsell, 1995), but all studies that have compared multiple stimuli inside the receptive field to a single stimulus inside the receptive field have found substantially larger attention effects in the former condition. These results are consistent with the proposal that attention operates to resolve ambiguities in neural coding that arise when multiple stimuli influence the response of a neuron.

Two questions are often asked in the context of these results. First, how is it possible for a neuron to “know” when multiple stimuli are present inside its receptive field? The answer to this question is not yet known, but we assume that the local circuitry within each visual area contains the information necessary to determine—at least coarsely—when a potentially ambiguous stimulus configuration is present. Second, why is attentional filtering limited only to a subset of neurons? The answer to this question is relatively straightforward: There is a cost to throwing away information. For example, it appears that neurons in area V4 achieve color constancy by integrating wavelength information across broad areas of the visual field, and widespread filtering within V4 might impair color constancy. Everyday experience also supports the proposal of limited filtering, because we do not experience the sort of “tunnel vision” that would result if most of the visual scene were filtered at a given moment. Thus, it is reasonable to suppose that attention is used sparingly and

FIGURE 47.4 Single-unit responses recorded from area V4 in the study of Luck and co-workers (1997a). (A) Responses of a single neuron to a blue vertical bar and to a green horizontal bar, presented individually inside the receptive field. (B) Responses of the same neuron when the two stimuli are presented simultaneously inside the receptive field and attention is directed to the blue vertical stimulus in some trial blocks and to the green horizontal stimulus in other blocks. (C) Average response across 29 neurons when attention is directed to an effective or an ineffective stimulus when both are presented simultaneously inside the receptive field. (D) Average response across the same neurons when only one of the two stimuli is located inside the receptive field. (Adapted with permission from Luck, Chelazzi, et al., 1997.)
that neural responses are suppressed only when neural coding becomes ambiguous.

The short-latency attention effects shown in figure 47.4 were obtained when attention was precued to a location before stimulus onset, and similar but delayed effects are found when the location of the to-be-attended object is not known in advance. In visual search tasks, for example, it appears that neurons in extrastriate cortex initially provide a coarse preattentive coding of the entire visual field so that a potential target can be located, followed by the focusing of attention onto this item for complete identification. This is illustrated in an experiment reported by Chelazzi and colleagues (1993), in which monkeys were presented with two items simultaneously and were required to determine if either of them was a target item. As shown in figure 47.5A, when one of the stimuli was an effective stimulus and the other was an ineffective stimulus for the neuron being recorded, the neuron initially gave a large response. Beginning around 175 ms poststimulus, however, the neuron's firing became suppressed if the target was the ineffective stimulus but continued unabated if the target was the effective stimulus. In other words, after an initial preattentive period—during which the neuron simply signaled the presence of an effective stimulus—the neuron's output began to reflect only the properties of the target stimulus. As in the case of precued attention, these effects were larger when both stimuli were inside the receptive field.

An analogous set of results has been obtained in ERP recordings from human subjects (Luck, Girelli, et al., 1997). In particular, modulations of an occipital ERP component called the N2pc wave (shown in figure 47.5B) closely resemble the monkey single-unit suppression effect shown in figure 47.5A. In addition to having a similar onset time, the N2pc wave is affected by a variety of experimental manipulations in the same manner as the single-unit attention effects observed by Chelazzi and his colleagues (Chelazzi and Desimone, 1994; Chelazzi et al., 1993). Specifically, the N2pc component is larger for conjunction discrimination tasks than for feature detection tasks, is larger when a target is surrounded by nearby distractor items, and is larger when a simple feature must be localized than when it must simply be detected. Thus, in both human and nonhuman primates, one major role of attention is to resolve the ambiguous neural coding that occurs when multiple items are presented within a neuron's receptive field.

**The Problem of Internally Generated Noise**

Despite the wide applicability of the ambiguity resolution theory, attention effects have also been observed under conditions that would not be expected to lead to ambiguous neural coding (Mangun and Hillyard, 1991;
Motter, 1993; Spitzer, Desimone, and Moran, 1988). For example, attended stimuli have sometimes been found to elicit larger responses than unattended stimuli even though these stimuli were presented sequentially and at widely separated locations (Mangun and Hillyard, 1988; Motter, 1993). Such effects have been observed primarily when the stimuli were difficult to identify or when a limited amount of time was available for perceiving or responding to the stimuli. To explain these modulations of sensory responses, several researchers have proposed that attention operates like a sensory gain control, amplifying attended channels compared to unattended channels (Hawkins et al., 1990; Hillyard, Vogel, and Luck, in press; Treisman, 1964). It is important to note, however, that increasing the gain of an input source will not always lead to an improvement in perceptual performance, because both the signal and the noise contained in the input will be amplified. A gain control mechanism can be useful, however, when perceptual performance is limited by the brain's internal noise rather than by noise in the input. As an analogy, imagine listening to a high-fidelity musical recording through headphones in a quiet room. At low volumes, one's own breathing and heartbeat may make it difficult to hear the music clearly, and turning up the volume even slightly may significantly improve the clarity with which the music is heard. In contrast, imagine listening to a very poor recording at a moderate volume level. Owing to the noise in the input source, the music may be very difficult to perceive clearly, but turning up the volume provides no improvement because an increase in gain increases the noise as well as the signal. Thus, an attention-controlled sensory gain control mechanism would be useful whenever performance is limited by internal noise rather than by noise in the input source.

There are many experimental tasks in which a simple gain control mechanism would be useful. For example, when an observer is required to discriminate between two similar colors, any intrinsic variability in spontaneous neural activity may impair discrimination performance, and increasing the gain at an early stage in the system would minimize the effects of the spontaneous activity. Similarly, when an observer is asked to make a very rapid response, a decision must be made before the visual system has acquired much information about the stimulus, and internal noise could lead to erroneous responses. In addition, high rates of stimulus presentation may increase internal noise levels because the neural responses elicited by irrelevant stimuli may contribute noise during the processing of relevant stimuli, and a sensory gain control mechanism may again prove useful under such conditions. These are exactly the conditions under which gain control–like modulations of sensory responses have been observed (e.g., Mangun and Hillyard, 1988; Mangun and Hillyard, 1991; Spitzer, Desimone, and Moran, 1988). Thus, it appears that attention plays at least two roles at the stage of perceptual processing—namely, resolving ambiguous neural coding by suppressing competing input sources and improving signal-to-noise ratios by controlling sensory gain.

**Conditions leading to late selection**

**Selection in Visual Working Memory** The term “working memory” refers to the temporary storage of information that can be manipulated according to task demands. Working memory has been divided into multiple subsystems, with a separate storage module for each sensory modality and a single central executive process that coordinates storage, manipulation, and retrieval for each module (see Baddeley, 1986). In this chapter, we concentrate on visual working memory, but similar principles apply to other types of working memory as well.

Working memory has at least two characteristics that make selective attention essential for its efficient operation. First, it is well known that the storage capacity of working memory is strictly limited, with a prototypical capacity of 7 ± 2 items in the domain of verbal working memory (Miller, 1956). The capacity of visual working memory is also quite small, with a limit of approximately 4 objects (Luck and Vogel, 1997). Consequently, it is important to store only the most relevant objects in working memory to avoid overloading its limited storage capacity. Second, the process of encoding an object into a durable form in working memory appears to be slow and resource-demanding. For example, Potter (1976) showed that subjects could identify pictures of complex, real-world scenes at very rapid rates of presentation (9 pictures per second), but they could not store the pictures in working memory at such rates. This result suggests that visual perception is typically much faster than working memory storage, and a selection process is therefore necessary to control which of the many perceived objects are stored in working memory.

Visual perception suffers from interference when many items are presented simultaneously at different locations, whereas working memory appears to be impaired when items are presented rapidly at a single location. These factors led us to propose that attention will begin to operate at the stage of perception in the former case and at the stage of working memory in the latter (Luck, Vogel, and Shapiro, 1996; Vogel, Luck, and Shapiro, in press). Evidence for this hypothesis has been derived from studies of the **attentional blink** phenomenon. When subjects view rapid sequences of stimuli, the
processing of T2, but they and others have more recently proposed that the attentional blink reflects an impairment at the later stage of working memory (Chun and Potter, 1995; Shapiro, Raymond, and Arnell, 1994). In other words, the visual system can identify every item in the stimulus stream, but subjects cannot store T2 in working memory while they are busy trying to store T1, and the perceptual representation of T2 is therefore overwritten by the subsequently presented stimuli (see Giesbrecht and Di Lollo, in press). This hypothesis was recently tested in a series of electrophysiological experiments in which T2-elicited ERPs were recorded in the attentional blink paradigm (Luck, Vogel, and Shapiro, 1996; Vogel, Luck, and Shapiro, in press).

The most informative experiment focused on the N400 component, which is highly sensitive to the degree of mismatch between a word and a previously established semantic context (Kutas and Hillyard, 1980). For example, a large N400 would be elicited by the last word in the sentence, "He put the towels in the washing machine and added a cup of POPCORN," but not by the last word in the sentence, "He put the towels in the washing machine and added a cup of DETERGENT." Similar results are obtained in experiments using sequentially presented word pairs: A larger N400 is elicited by the second word if it deviates from the semantic context established by the first word [e.g., a larger N400 is elicited by the second word in lime-NICKEL than in lime-LEMON (see Kutas and Hillyard, 1989)]. Because a semantic context violation cannot be detected before the word's meaning has been assessed, the presence of a larger N400 component for a mismatching word indicates that the word has been identified to the point of lexical/semantic access. Therefore, the presence of a normal N400 component during the attentional blink period would provide strong evidence that words presented during the attentional blink are fully identified, even though the observer cannot accurately report them.

This prediction was tested in an attentional blink experiment in which T2 was a word that either matched or mismatched a semantic context established at the beginning of each trial. The results of this experiment are summarized in figure 427. Consistent with previous attentional blink experiments, T2 accuracy was impaired when T2 was the third item following T1 (lag 3) compared to when T2 was the first or seventh item following T1. In contrast, the N400 component was not suppressed during the attentional blink period, which indicates that the words were identified to the point of lexical/semantic access during this period, even though behavioral discrimination accuracy was substantially impaired. To demonstrate that the N400 component is actually sensitive to perceptual
impairments, an additional experiment was conducted in which varying amounts of visual noise were added to the T2 stimulus, making it more difficult to perceive. This manipulation had a substantial influence on N400 amplitude, demonstrating that the N400 is quite sensitive to changes in perception. These results therefore provide strong evidence that attention operates at a postperceptual stage when stimuli are presented in a rapid sequence at a single location rather than being presented simultaneously at multiple locations.

To determine more precisely the stage at which processing is impaired during the attentional blink, an additional experiment was conducted to examine the P3 wave, which occurs after the N400 component and has been proposed to reflect the updating of working memory (Donchin, 1981). Unlike the N400 component, the P3 wave was found to be completely suppressed during the attentional blink period. This provides additional evidence for a role of attention in limiting the storage of information in working memory.

Selection during response processing ERP recordings have also been used to examine even later stages of processing, such as the response selection processes that may be subject to attentional control when multiple responses must be made in a short period of time. For several decades, cognitive psychologists have studied this type of dual-task interference by means of the psychological refractory period (PRP) paradigm, which is illustrated in figure 47.8. In a typical PRP experiment, two targets (T1 and T2) are presented on a given trial, and the subjects is required to make a separate speeded response to each target. When the SOA be-

![Figure 47.7](image-url)  
Figure 47.7 (A) N400 amplitude and T2 accuracy from the study of Vogel, Luck, and Shapiro (1998). When T2 was the third item after T1, T2 accuracy dropped dramatically, but there was no suppression of the N400 component. (B) Results from an additional experiment in which simultaneous visual noise was added to T2 and subjects were not required to report T1. As more noise was added, accuracy and N400 amplitude both declined, indicating that the N400 component is sensitive to manipulations of perceptual quality. (Data from Vogel, Luck, and Shapiro, in press. Copyright © 1998 by the American Psychological Association. Adapted with permission.)

![Figure 47.8](image-url)  
Figure 47.8 Paradigm (A) and results (B) from a typical PRP experiment. On each trial, two targets are presented (T1 and T2), separated by a variable SOA. Independent responses are made to the two targets (R1 and R2), resulting in two reaction times on each trial (RT1 and RT2).
responses according to the rules of the task). To distinguish between these two possibilities, it is possible to examine the latency of the P3 wave, which can differentiate between stimulus categorization and response selection.

When a subject is instructed to discriminate between two or more stimulus categories, the P3 wave is typically larger for relatively improbable stimuli than for relatively probable stimuli (see review by Johnson, 1986). As an example, consider a task in which subjects view a sequence of names and must indicate whether each name is used for males or females (e.g., John vs. Sue). If each name is presented once, but 90% of the total are male names and 10% are female names, then a larger P3 wave will be elicited by the female names. The P3 wave is thus sensitive to the probability of a task-defined category, not the probability of an individual stimulus. Accordingly, an effect of probability on P3 amplitude cannot occur until the subject has determined the task-defined category of the stimulus, and the latency of the P3 wave is therefore highly sensitive to the amount of time required to perceive and categorize a stimulus.

To test the hypothesis that the slowing of responses at short SOAs in the PRP paradigm reflects a delay in response selection, the P3 wave was examined in the paradigm illustrated in figure 47.9A (Luck, 1998). T1 was an outline square, and the subjects were required to press one of two buttons with one hand to indicate whether the square was red or green. T2 was a letter, and the subjects pressed one of two buttons with the other hand to indicate whether it was an X or an O. To elicit the probability-sensitive P3 wave, one of the two T2 alternatives was presented less frequently than the other (25% versus 75% of trials). The P3 wave was then isolated by constructing infrequent-minus-frequent difference waves. This subtraction eliminated any ERP activity that was unrelated to the probability of T2, including T1-elicited ERP activity. If the slowing of the response to T2 at short SOAs in the PRP paradigm is caused by interference in perception or categorization, then the latency of the P3 wave should also be delayed at short SOAs. In contrast, if the slowed RTs are caused by interference with response selection, then the P3 wave should not be delayed.

The results of this experiment are summarized in figure 47.9B. As in previous PRP experiments, the RT for T2 was significantly slowed at short SOAs. In contrast, P3 latency was only slightly delayed at short SOAs, which indicates that the slowing in RT reflects interference at a stage that follows perception and categorization, presumably response selection. A follow-up experiment showed that manipulations of T2 luminance produced large changes in P3 latency, demonstrating the sensitivity of the P3 wave to manipulations of stimulus evaluation processes. Thus, it appears that the slowing of reaction time in the PRP paradigm is due primarily to delays in response selection, which is consistent with the model of Pashler (1994).

Conclusions

For several decades, cognitive psychologists have debated whether attention operates at an early stage or at a late stage, but it is now becoming clear that attention operates at both early and late stages (and at intermediate stages, as well). Moreover, attention researchers are now beginning to understand the specific reasons why atten-
tion may be necessary at a given stage under a given set of conditions. In this chapter, we have identified two computational problems—the binding problem and the problem of internally generated noise—that provide good explanations of why attention is necessary at the stage of visual perception under some conditions but not under others. We have also pointed out two key limitations at the stage of working memory—the limited storage capacity and the slowness of the storage process. Presumably, attention will operate at the stage of working memory whenever these two limitations may impair task performance. Finally, we have provided evidence that attention may sometimes be required at the stage of response selection, although it is not yet clear why it is not possible to select different responses in parallel. It seems likely that attention operates at other stages of processing in addition to those discussed here and that the coarsely defined stages described in this chapter can be subdivided more precisely. A major challenge for the future is to determine how these multiple attentional mechanisms operate in a coordinated manner to maintain unity of behavior (Posner and Peterson, 1990).

ACKNOWLEDGMENTS Preparation of this chapter was supported by grants from the National Institute of Mental Health (MH-56877-01 and MH-25594-24), the Human Frontier Science Program (RG0136), and the Office of Naval Research (N00014-93-1-0942).

REFERENCES


