CHAPTER FOUR

Attentional Limitations in Dual-task Performance

Harold Pashler
University of California, USA

James C. Johnston
NASA Ames Research Center, Moffett Field, USA

INTRODUCTION

People's ability (or inability) to do different activities or tasks at the same time is a topic of much interest not only to psychologists, but also to the proverbial "person in the street". It is natural to wonder about what we as human beings can and cannot do. An understanding of our limitations should also have practical value, because the intelligent design of human/machine systems depends as much on knowing the capabilities of people as it does on knowing the capabilities of machines. Human performance limits have played an important role in catastrophes that have occurred in aviation and other fields; a better understanding of those limits might help in designing systems and procedures that can minimize the frequency of such disasters.

Simultaneous performance of different tasks is intellectually intriguing as well. The limitations on simultaneous cognition may provide important clues to the architecture of the human mind. The notion that dual-task performance limitations have implications about the "unity of the mind" occurred to people long before the present era of information-processing psychology. In the late nineteenth century, for example, the educated public was fascinated with a phenomenon called "automatic writing", in which People were claimed to be able to write prose while carrying out other tasks (see Koutstaal, 1992).

This chapter provides an overview of research on attentional limitations dual-task performance. The organization of the chapter follows a plan

that is typical in present-day psychology: we begin with mental processes at the "front end" (i.e. analysis of incoming sensory stimuli), and then proceed to more "central" processes (e.g. decision-making, memory storage, memory retrieval, and action planning).

What is meant by "attentional limitations"? A formal definition of the term "attention" is not presently available, nor is it likely that a compelling definition can be arrived at from *a priori* considerations alone. However, it seems sensible to start with at least some rough characterization of what sort of phenomena we describe as attentional. In this spirit, we offer two criteria that a performance limitation must satisfy to be called attentional. First, the limitation must not be a direct consequence of the structure of the human body or its sensory or motoric apparatus. By this criterion, our inability to drink a cup of coffee and type at the same time is not attentional (at least not exclusively attentional), nor is our inability to read a word in the newspaper 10 degrees off fixation (as this is a consequence of the low density of photoreceptors in the periphery of the retina). Second, an inability to perform two tasks at the same time to a given criterion of performance is attentional only if a person could voluntarily perform either task alone to that criterion under essentially the same conditions. By this criterion, our inability to comprehend two spoken messages at the same time probably does qualify as attentional, while our inability to read two superimposed visual words (one masking the other) probably would not.

Roughly speaking, then, attentional limits are caused by limitations on those parts of mental machinery or processes that are normally subject to voluntary control or direction. We can use this definition without assuming (as ordinary language seems to assume) that that these limitations reflect a single underlying entity or process called Attention. From a scientific point of view, that assumption is probably best regarded as mere folklore; attentional limitations may turn out to reflect a great variety of mechanisms or processes.

Overview of Theories of Dual-task Interference

Before turning to empirical findings, we begin by considering a few theoretical ideas and concepts that have been widely applied in trying to understand attentional limitations. The first concept is that of a strict processing *bottleneck.* This refers to the idea that certain critical mental operations are carried out sequentially, and must be carried out sequentially. When this limitation applies, a bottleneck arises whenever, in a dual-task situation, two tasks require a critical mental operation at the same point in time. This kind of account is generally referred to as a bottleneck or single-channel model. The most obvious explanation for the existence of bottlenecks, if they do in fact exist, would be that the mind/brain contains

only a single device or mechanism capable of carrying out the operation(s) in question. Other interpretations are possible, however. For instance, two operations that are carried out in different neural machinery might inhibit each other, thereby making it possible for only one or the other to operate at any given time. Naturally, there could be not just one, but two or more distinct bottlenecks associated with different types of mental operations, and bottlenecks might depend not only on the type of mental operation to be performed but also on the types of material to be processed and the extent to which the operation had been practiced.

Many theorists have argued for a less discrete analysis of dual-task performance limitations. They have suggested that there may be one or more pools of processing "resources" (sometimes equated with "effort" or "mental fuel") that can be divided up among different tasks or stimuli in a graded fashion. That is, when more processing resources are devoted to one task or stimulus this leaves a little less for others. On this account, processing for different tasks proceeds in parallel but the rate or efficiency of the processing depends on the capacity available to the task (among other factors). This conception will be referred to as *capacity sharing.* The single-channel bottleneck and capacity-sharing models provide very different pictures of our mental machinery; in the single-channel conception, certain aspects of mental processing are invariably sequential, while in the capacity view, processing on different tasks is simultaneous but occurs more slowly due to the reduction in available resources.

A third interpretation of attentional limitations attributes interference to *crosstalk* or other impairment in performance that hinges directly on the specific content of the information being processed. It has long been suggested that when two tasks are more similar, performing them together causes more interference than would be the case with very different tasks (e.g. Paulhan, cited by James, 1890). Some theorists have suggested that even when there is adequate machinery to carry out different tasks at once, keeping processing streams separate may be an important cause of dual-task interference. This predicts that the interference depends on the similarity or confusability of the mental representations involved in each task (e.g. Navon & Miller, 1987). This kind of theory is not entirely incompatible with the idea of a bottleneck. One could suppose, for example, that certain mental operations operate sequentially precisely because if they were allowed to run concurrently, crosstalk would occur (see Kinsbourne, 1981, for suggestions along these lines).

The three broad approaches sketched here do not exhaust the space of possible theories of dual-task performance. Further alternatives can be considered, and various hybrids can be constructed out of the models just mentioned. However, the concepts of capacity, bottleneck, and crosstalk provide an adequate framework for appreciating the research described in

the remainder of this chapter and we will defer consideration of more complicated models until the empirical motivation for complications has emerged.

EMPIRICAL EVIDENCE ON LIMITATIONS

We turn now to the empirical evidence on human dual-task performance limitations, beginning with sensory and perceptual analysis and moving from there to more central processes. This review is necessarily incomplete but fairly representative of a large and growing literature.

Perceptual Processing Limitations

As mentioned in the Introduction, the earliest information-processing theories of attention tended to assume very severe limitations in perceptual processing. Broadbent's Filter Theory, for example, contended that people are unable to identify more than a single spoken word at one time. In the late 1960s and early 1970s, several theorists proposed a radical departure from this approach, arguing that sensory and perceptual processes are subject to no attentional limitations whatever. This new view was termed the "late-selection theory of attention" (referring to the idea that selection occurs only after all stimuli are identified unselectively). At first glance, this notion might strike the reader as absurd: surely an organism's capacity to do anything must be limited. It should be kept in mind, however, that attentional limits are only one factor potentially limiting performance. Claiming that attentional limitations on perception do not exist does not, therefore, imply that perceptual systems do a "perfect" job of analyzing a stimulus or that people can recognize an unlimited number of objects at the same time; it merely implies that these systems analyze any one stimulus just as effectively whether other stimuli are being processed at the same time or not.

Whether plausible or not, the late-selection account rested from its inception on a rather thin evidence base from its inception. The main support for the theory consisted of demonstrations that, even when people try to ignore certain stimuli (e.g. letters to the left or right of fixation, words spoken to an unattended ear), these stimuli are nonetheless analyzed semantically, at least to some extent (e.g. Eriksen & Hoffman, 1973). Naturally, the fact that to-be-ignored stimuli are sometimes analyzed does not imply that there are no attentional capacity limitations, especially when the evidence of unwanted processing involves only a few simple stimuli like letters.

More direct empirical assessments of capacity limitations in perceptual processing involve "divided attention" tasks in which people try to identify a number of objects simultaneously. The most obvious task to use for this purpose is one that requires observers to report all the stimuli they can (e.g. to read off a display of briefly exposed letters). This "whole report task" was

first studied in detail by Sperling (1960). Sperling's findings disclosed that limitations in storing or retaining stimuli in short-term memory often prevent stimuli from being reported even when conditions would allow them all to be identified successfully (see also Estes & Taylor, 1964). To assess perceptual capacities apart from memory limitations, one needs a task that allows subjects to demonstrate what they have identified without having to hold much information in memory.

For this reason, most contemporary research on visual processing capacity limitations has relied on detection or search tasks (Estes & Taylor, 1965). In these experiments, people report the presence or absence of a prespecified target somewhere in a search display or, in some cases, choose which of several alternative targets was present. As Wolfe describes in Chapter 1, when the number of distractors in search displays is increased, response times generally increase too. This increase is much greater when the target and distractor differ only in some subtle or complex fashion, e.g. when they consist of the same parts but arranged differently (e.g. Logan, 1994) or when targets and distractors vary along some continuous dimension like length or orientation and the difference between them is subtle (e.g. Treisman, 1991). This finding strongly suggests the existence of capacity limits in perceptual analysis. For technical reasons, however, the inference is not absolutely secure. For purely statistical reasons, accuracy falls as the number of items to be searched is increased (basically because each distractor represents an additional opportunity for a "false alarm"). Increases in response times could potentially occur because people compensate for this by taking longer with larger display set sizes, not because capacity limitations arise (Duncan, 1980a; Palmer, 1994).

Detection studies that focus on the accuracy rather than speed of visual search performance have also been carried out, usually using objects briefly exposed and followed by pattern masks; pattern masks serve to curtail visual persistence and ensure that visual analysis takes place immediately rather than at the subject's convenience. One particularly incisive experiment from the point of view of assessing capacity limitations was carried out by Shiffrin and Gardner (1972). They required observers to search a display of four alphanumeric characters for a target character (each item followed by a mask). In the Simultaneous Condition, all four characters were flashed at the same time. In the Successive Condition, they were flashed one or two at a time, with pauses in between flashes. If capacity limitations were at work, the successive condition should provide an important advantage: each item can benefit from more capacity than it could in the simultaneous condition, where capacity must be divided up among all four items. What Shiffrin and Gardner observed, however, was essentially no difference in accuracy between the two conditions. This striking finding has been replicated by a number of other investigators (e.g. Duncan, 1980b), and argues strongly

that at least four characters can be processed in parallel without attentional limitations.

Although the result is consistent with the strong claims of late-selection theories, further experimentation using the same technique showed that capacity limitations arise whenever processing load is increased beyond a certain point. When the target/background discrimination is more difficult or larger display set sizes are used, for example, accuracy in the simultaneous condition is often substantially worse than in the successive condition (Duncan, 1987; Kleiss & Lane, 1986). This finding fits well with the results of speeded search tasks (see Wolfe's Chapter 1) and refutes the late-selection theory discussed earlier.

What causes processing limitations in perceptual analysis? Unfortunately, at present we have few clues to help in answering this interesting question. Recent studies of the time course of interference, and of the correlation between performance on simultaneous discrimination tasks, suggests that capacity overload may typically produce graded capacity sharing rather than sequential processing (Duncan, Ward, & Shapiro, 1994; Miller & Bonnel, 1994). There is no evidence that performance limitations result from crosstalk in processing multiple items, because the interference seems to be comparable whether the tasks involve similar or different sorts of judgments (Duncan, 1993). However, interference is clearly worse when inputs involve a single sensory modality (e.g. audition or vision) compared to inputs presented in different modalities (Treisman & Davies, 1973).

These observations suggest several tentative conclusions. Identifying stimuli probably requires processing resources that are specific to a particular sensory modality. Particular perceptual discriminations require different amounts of processing capacity, depending on the difficulty of a given discrimination. Different objects can be analyzed in parallel and independently, but only so long as total capacity demands are not exceeded. When they are exceeded, perceptual analysis becomes less efficient, perhaps typically resulting in parallel processing at reduced efficiency.

How do these capacity limits affect our perception of the sights and sounds of less austere displays such as those we usually encounter in daily life? Here our knowledge is sadly limited. Studies in which observers look at rapid sequences of pictures (many frames presented per second in the same part of the visual field) reveal that people can comprehend a scene very rapidly (subject, of course, to acuity limits of the periphery). For example, observers can fairly reliably detect an object out of place in scenes viewed at rates of approximately 100-200 msec per picture (Biederman, Teitelbaum, & Mezzanotte, 1983). The fact that processing is rapid does not imply that it is free of capacity limitations, of course; any given object in a scene may be identified more slowly than it would be in isolation. It seems reasonable to suspect that this is so, based on the laboratory studies described earlier, but there are so many differences between perception of scenes and visual search involving letters and symbols that this is only a conjecture.

Central Processing Limitations

What sort of processing limitations arise at central levels of the cognitive system, i.e. in the neural/mental machinery responsible for thinking, decision making, and planning actions? To isolate limitations in central processing, we need experimental methods that do not overload perceptual limitations. One simple precaution is to use tasks whose perceptual requirements are relatively slight. Given the results described earlier, another natural precaution is to use different input modalities in each task (typically vision and audition), making it less likely that any single perceptual mechanism will be overloaded. These precautions are sensible, but not definitive; the locus of any particular dual-task interference cannot be assumed *a priori.*

Before turning to the empirical literature, it is worth pausing momentarily to ask what ordinary experience might teach us, if anything. Most of us have reflected on the fact that we sometimes seem able to perform two daily tasks at the same time, e.g. carrying on a conversation and driving a car. At some crude level of description, this is certainly possible: the car does not end up in a ditch and our friends are (usually) not offended. What does this tell us about the extent of concurrent mental processing? Although the car is certainly moving while the driver is speaking, often it would continue on a perfectly reasonable trajectory for several seconds without any action whatever on the part of the driver. As for the conversation, even under normal circumstances, conversing involves intermittent behavior and partially redundant messages. Therefore, the brute fact that people drive and converse does not necessarily indicate that parallel processing of all aspects of language processing and driving is possible. Our casual observations would be equally consistent with the possibility that our brains do what single-processor computers often do to juggle multiple tasks or users: work on one task at a time, but alternate rapidly between them in order to respond to inputs in a timely fashion.

Computer time-sharing normally requires that the computer must be able to "buffer" (hold in temporary memory) a considerable amount of information. Computers usually have buffers both for inputs that have not yet been fully processed and for outputs that have been planned but not yet carried out, as well as various internal buffers. As it happens, human beings are also equipped with a variety of memory buffers that seem roughly suitable for such a "buffer and switch" processing strategy (Baddeley, 1986).

Ordinary experience, then, provides little insight as to which mental operations can occur at the same instant; laboratory studies are plainly

needed to sort this out. Research in the lab has confirmed the casual observation that people can sometimes perform two continuous tasks concurrently with only modest loss in performance, when these tasks involve no obvious conflicts in input or output modality. Examples include playing the piano and shadowing spoken words (Allport, Antonis, & Reynolds, 1972) and, for a skilled typist, typing a manuscript while shadowing (Shaffer, 1975). Even more remarkably, Spelke, Hirst, and Neisser (1976) were able to train subjects to take dictation (speech input, manual output) while reading aloud (visual input, vocal output). In each of these cases, dual-task performance was very good, although error rates were usually somewhat higher than in the single-task conditions.

Although tasks like typing and reading aloud might seem to require more continuous cognitive activity than driving of conversing, smooth combination of these tasks might still reflect a switching strategy. One factor that might facilitate switching is sometimes termed "chunking": after people have practiced a task like typing, they seem able to plan and execute larger and larger response units. For a skilled typist, for example, the primary or highest-level unit at which actions are planned is probably the word rather than the individual letter. Several observations support this idea. For example, the "eye-hand span" (the lag between which letter a typist fixates and which letter they are typing at the same instant) is considerable, and grows longer as a typist acquires expertise (Salthouse & Saults, 1987).

Suppose, then, that seemingly continuous activities like typing actually involve discontinuous planning of relatively large output "chunks". In that case, if people can smoothly perform two such tasks concurrently, this may be because the planning operations are occurring at different moments, not because the two tasks are performed completely independently. Consider the analogy of a person doing the laundry and cooking dinner. Although few people can load the laundry and stir a frying pan at the same instant, for example, nonetheless one can sometimes schedule things so that the dinner and the laundry are both completed at roughly the same time that they would be if carried out by themselves. At a finer time-scale, the same may be true for the central operations involved in tasks like typing and reading aloud.

In the light of these possibilities, to determine whether people can carry out different central mental operations independently, we need to look at the speed with which they can respond to individual stimuli presented close together in time. Buffering and switching may suffice to produce continuous behavioral streams but the lag between individual stimuli and the corresponding responses should nonetheless provide telltale indications of delay. To detect these signs, we need to use experimental designs in which stimuli are presented at discrete moments in time, and to determine the latency of responses that are unambiguously related to particular stimuli.

A rather austere type of experiment fits these requirements. This design involves discrete trials; the subject must, on each trial, perform two tasks, responding to each of two different stimuli presented in rapid succession (S1 and S2). The interval between the onset of the two stimuli (known as the stimulus onset asynchrony, or SOA interval) is varied, typically from short intervals of 50 ms or so up to intervals of half a second or longer. This kind of experiment was apparently first tried by Telford (1931), and since then hundreds of studies have confirmed a very basic result: as the SOA between two stimuli is shortened, responses to the second stimulus are delayed, often by hundred of milliseconds. This finding is observed with almost all tasks involving choice (i.e. response uncertainty), although some exceptions and potential exceptions will be discussed later.

Why should there be such a delay? A natural interpretation is that subjects are unable to carry out certain processing involving the second stimulus until they have finished with the first. When the effect was first observed, some thought it was analogous to the refractory period of neurons—after producing a neural spike, neurons are temporarily inhibited from firing again for a very brief period. Although the delays in the behavioral case are usually much longer than the neural refractory period, and although the two phenomena differ in other important respects, the delay in responding to the second of two stimuli has been christened the Psychological Refractory Period (or PRP) effect, and for better or worse, the term has stuck. We will use the term PRP Paradigm to refer to the method, leaving open the question of whether the analogy to neural refractoriness is illuminating or not.

Let us look more closely at an actual PRP experiment. For convenience we will use a study of our own (Pashler & Johnston, 1989, Experiment 1). The experiment was carried out with a microcomputer. Subjects were required to perform two completely unrelated choice response-time tasks. For Task 1, the stimuli were a 300 Hz tone and a 900 Hz tone. Subjects responded by pressing one of two adjacent keys on a keyboard with fingers of the left hand. For Task 2, the stimuli were the letters "A", "B", or "C"; subjects responded by pressing one of three adjacent keys with fingers of the right hand. Each trial began with subjects fixating a mark at the center of the CRT screen. After 1000 ms this mark was extinguished, and after another 200 ms one of the two possible tones for Task 1 sounded for 33 ms. After a variable SOA (50, 100, or 400 ms) one of three Task 2 stimuli appeared on the CRT screen. Subjects were instructed to respond rapidly and accurately on both tasks, but the instructions particularly emphasized the importance of responding rapidly on Task 1. In a control condition, subjects were asked to perform only Task 1; Task 2 stimuli were presented, but subjects did not respond to them.

Figure 4.1 shows results from the experiment. The solid curve shows mean RT for Task 1 in the dual-task condition. It was about 600 ms

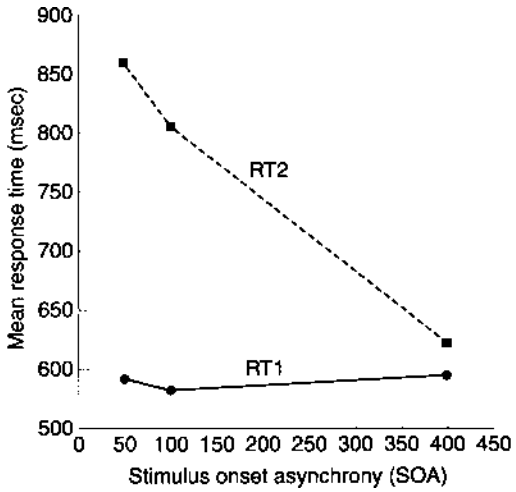

FIG. 4.1. RT for Task 1 and Task 2 in a dual-task experiment requiring separate responses to a tone (first stimulus) and a letter (second stimulus). Stimulus onset asynchrony refers to the time between onset of tone and letter. RT1 is time for response to tone; RT2 is time for response to letter. (Data from Experiment 1, Pashler & Johnston, 1989.) The slowing of RT2 at shorter SOAs is referred to as the psychological refractory period (PRP) effect.

regardless of the SOA delay in presenting the Task 2 stimulus. The dashed curve shows RT2 in the dual-task condition. Unlike many functions, this one is easiest to think about starting from the right-hand side. At the longest SOA (400 ms), when the two tasks could generally be done sequentially, RT2 was a little over 600 ms. But at the shortest SOA (50 ms), when both stimuli need to be processed at almost the same time, RT2 was several hundred ms higher—about 850 ms. The fact that subjects are delayed several hundred ms in making the response to that letter stimulus (the PRP effect) is what we would expect from the "buffer and switch" strategy discussed earlier. Note that the dual-task interference found here is clearly evident only in RT data; the data show no increase in errors on Task 2 at short SOAs (in fact subjects made more errors at the longest SOA).

Although RT1 in the dual-task condition was little influenced by SOA, it was consistently slower than the response time for the same task performed alone in a control condition. This is a typical finding, and many theorists attribute it to the difficulty in attaining an optimal level of preparation for both tasks (cf. Gottsdanker, 1980; Pashler, 1994). If this is correct, we should expect slowing even if Task 2 were omitted altogether on some trials, and this has been observed as well (Ruthruff & Pashler, submitted).

As mentioned earlier, the basic PRP result—a dramatic slowing in RT for the second task, when stimuli for both tasks are presented in rapid succession—has been replicated many times, with a variety of different tasks (for reviews of early work, see Bertelson, 1966, and Smith, 1967). These results have suggested to many investigators that the mind may, in certain respects at least, contain a "single-channel" information processor, capable of working on only one task at a time. The simplest version of such an account would claim that all aspects of the second task are delayed until all aspects of the first task are finished. The data in Fig. 4.1 show, however, that both tusks are often completed in less time than the sum of the time it would for each task to be performed in isolation. (The average time elapsed from S1 to R2 was 909 ms at the shortest SOA, while the times to perform each task alone were roughly 590 and 500 ms.) At a very crude level, then, there are signs of overlapping processing. This suggests that if there is a bottleneck, it probably involves some but not all of the processes involved in carrying out the two tasks. The earliest proponent of a bottleneck, Welford (1952, 1980), suggested that single-channel processing occurred in the central stages of each task—stages that he termed "stimulus-response translation". This *central bottleneck model* is at least grossly consistent with the data from the experiment we have been looking at.[1]

Recent evidence provides more definitive evidence about the validity of the single-channel hypothesis. Two separate issues can be distinguished: first, whether PRP interference is attributable to a bottleneck at all; and second, if it is, what the functional locus of this bottleneck might be.

Is PRP Interference Caused by a Bottleneck?

Although many early researchers favored a bottleneck explanation for PRP interference, others advocated a capacity-sharing analysis. For example, in his book *Attention and Effort,* Kahneman (1973) proposed that PRP interference typically reflects a slowdown in processing that occurs when general processing capacity is shared between tasks at short SOAs. Assuming (as seems reasonable) that reduced capacity results in slower Processing, this readily accounts for the basic PRP effect.

Several further pieces of evidence have sometimes been taken to support a capacity account. Kahneman noted that in many PRP studies, responses were slower not only on Task 2, but also on Task 1. Capacity theory, because it treats Task 1 and Task 2 symmetrically, has a natural account for RT1 slowing when it occurs. As capacity is divided between both tasks, processing for each task should operate at a slower than normal rate, resulting in delays of both responses. Bottleneck theory itself provides no account of RT1 slowing, but it can be amended to do so in a reasonable fashion. For one thing, subjects given no instruction to keep the first task response as fast

[1]He also suggested that monitoring the response occupied the single channel as well.

as possible might sometimes switch to working on Task 2 before finishing Task 1. For another, subjects may often "group" the two responses, i.e. withhold the first response until the second response has been selected. Both of these strategies are likely to be under some degree of conscious control, and indeed, when subjects are told to produce R1 as fast as possible (as in the experiment described earlier), there is little sign of R1 slowing.

Although capacity theories have not been shown to make particularly distinctive predictions,[2] the central bottleneck model does make very detailed predictions. Fig. 4.2 shows the model graphically, with time running from left to right and the upper three connected boxes representing Task 1 and the lower three boxes representing Task 2. In the spirit of Sternberg's (1969) stage analysis, it is assumed that processing on each task can be divided into three general stages, with each stage normally commencing as soon as the preceding stage is finished. We will discuss later and in more detail what might be accomplished in each stage, but for now one can think of stage A as "early" stimulus processing, stage B as central processing, and stage C as "late" response-related processing (each of these could be subdivided into further stages, naturally, but for present purposes this is not necessary). The assumption of a central bottleneck can be stated slightly more formally, as follows: (1) any stage A or stage C can proceed on each task regardless of what is happening on the other tasks, but (2) stage B can operate for only one task at a time. If stage 1B is running, stage 2B cannot run and vice versa. From this, one can derive equations for both RTs:

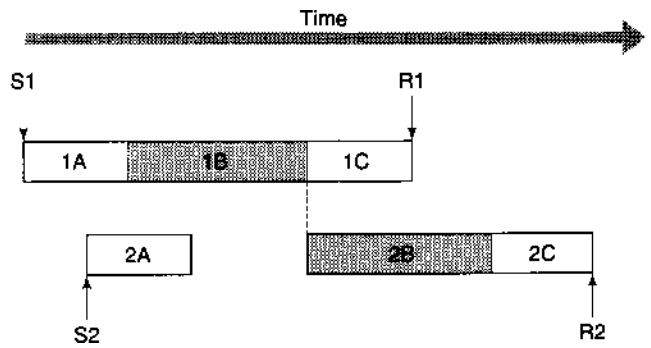

FIG. 4.2. Sequence of processing stages hypothesized in most general versions of the central bottleneck model of PRP effect. Stages 1A, 1B, and 1C comprise Task 1, while 2A, 2B, and 2C comprise Task 2. A fundamental constraint is that processing in the shaded stages cannot operate simultaneously; hence, a bottleneck arises when S1 and S2 are presented close together in time, as they are here; stage 2B cannot begin until stage1B has been completed.

[2] But see McLeod (1977) for an example one highly restricted version of capacity-sharing theory that might make stronger predictions.

(i) RT1 - 1A + 1B + 1C
(ii) RT2 = Max(1A + 1B + SW,SOA + 2A) + 2B + 2C-SOA

Note that RT1 is simply the sum of the component stage times for Task 1 However for Task 2, the central stage 2B cannot begin until both the resources required by the central stage are released (determined by 1A + 1B

SW, where SW is any switching time required to move central resources from Task 1 to Task 2; in the figure SW = 0) *and* early stimulus processing in Task 2 is finished, providing the information input to stage 2B, (determined in SOA + SW). Because of the "and" relation, RT2 on any given trial is determined by whichever of these two terms is greater.

In exploring this model, let us begin with the (surely unrealistic) simplifying assumption that stages times are deterministic. It then follows that a graph of RT2 against SOA should look like Fig. 4.3 [elbow curve]. To the right of the elbow, at longer SOAs, there is never a wait for the central resources to be released from Task 1, so SOA has no effect on RT2. At short SOAs, RT2 is always determined only by how rapidly stage 2B finishes. The time at which S2 is presented no longer affects when R2 occurs, so the combined interval from the beginning of S1 to R2—which is equal to SOA

+ A2—is a constant. This means that for every one millisecond increase in SOA, RT2 decreases by one millisecond. Thus we arrive at the very specific prediction that the graph of RT2 against SOA should have a left segment with a slope of -1, a right segment with a slope of 0, and a bend at the point where the processor and the results of stage 2A become available simultaneously.

If we allow stage durations to vary stochastically, the graph of the mean RT2 against SOA should look more like the dotted curve: the elbow is now

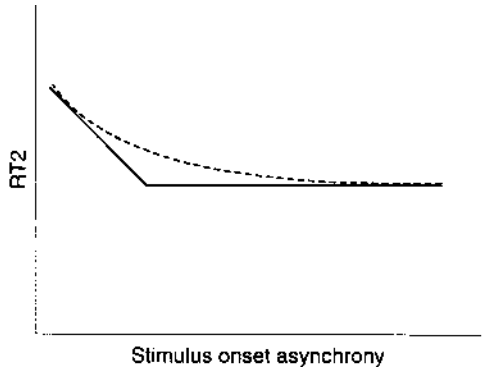

FIG. 4.3. Shape of function relating RT2 (vertical axis) to SOA (horizontal axis), as predicted by a deterministic version of the bottleneck model (solid line) and a more realistic version that assumes stochastic variation in stages durations (approximated with dotted line).

smoothed out by trial-to-trial variation. There may be some portion of the curve on the left with a slope of -1, but only if there is range of SOAs over which stage 2B always waits on completion of central stages in Task 1. There will be some range of SOAs on the right of the curve over which the slope will be 0 if, over this range, stage 2B never waits on completion of the central stages in Task 1.

One clear prediction of this model is that a graph of RT2 against SOA should be declining but positively accelerated. This is virtually always found. The prediction that a -1 slope occurs at very short SOAs is often at least approximately true (the slope is -1.06 for the Pashler & Johnston data shown in Fig. 4.1) and sometimes very precisely true. In an experiment of McCann and Johnston (1989) to be discussed shortly, where an unusually large amount of data was obtained by testing the same subjects over six days, four different curves of this kind had measured slopes between SOA 50 and SOA 150 of -1.02, -1.00, -1.00, and -.97. Note that if Task 1 is relatively fast and/or stage 2A requires unusually lengthy processing, then -1 slopes may not be obtained. Note also that the prediction does not hold if subject sometimes take time out from Task 1 to do some processing on Task 2, or use a grouping strategy.

A more realistic version of the model assumes that stage times have stochastic variability. This makes predictions about how RT1 and RT2 covary across trials. Generally speaking, anything that causes stage 1B to finish later (anything slowing stages 1A or 1B) will "push on" RT2 as well as RT1. This leads to the prediction that RT1 and RT2 will be positively correlated. High positive correlations are indeed observed at short SOAs (Gottsdanker & Way, 1966; Pashler & Johnston, 1989). More recently, it has often been noted that at long SOAs, RT2 tends to be influenced only by unusually long RT1s, whereas at shorter SOAs, it is influenced by progressively shorter and shorter RT1s (e.g. Pashler, 1989; Pashler & O'Brien, 1993).

The prediction that longer times for stages 1A and 1B will "push" onto RT2 at short SOAs but not at long SOAs also holds if these stages in Task 1 are lengthened by experimental manipulations rather than spontaneous variability. McCann and Johnston (1989) manipulated the difficulty of stimulus processing on Task 1, using a factor that increased RT1 by 29 ms. The effect of this Task 1 difficulty factor on RT2 was 30 ms at SOA 50, 28 ms at SOA 150, 16 ms at SOA 300, and only 2 ms at SOA 800. (These results are an average over two different types of Task 1, one auditory and one visual; virtually the same pattern was observed in each case.) This is just the pattern predicted by central bottleneck theory. Although capacity theory can predict effects of the difficulty of one task on performance in another, a millisecond for millisecond "propagation" of Task 1 effects onto RT2 at short SOAs is a hallmark of a true bottleneck.

Perhaps the most striking predictions from central bottleneck theory arise when difficulty of Task 2 is manipulated, however. Suppose that, by means of some difficulty manipulation, we increase the duration of some processing stage on Task 2 by *k* ms. What does the model predict? The predictions differ, depending on whether the manipulated stage occurs before or after the "gap" in the timeline for Task 2. This gap in the timeline represents the postponement of stage B of Task 2 until the completion of stage B in Task 1. The top panel of Fig. 4.4 shows what the model predicts if we manipulate the duration of stage 2B. As the change is after the "gap" is over, the consequence is simply that RT2 increases by the same amount *k* by which stage 2B was lengthened. Hence the prediction is that RT2 will be increased by *k* regardless of SOA. Thus, the effect of the stage manipulation will be additive with the effect of SOA on RT2.

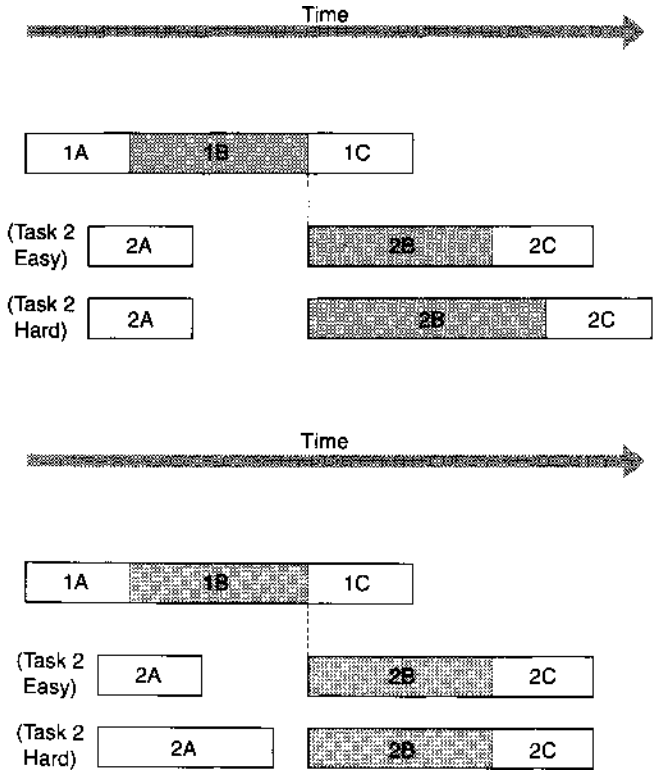

FIG. 4.4. Effects of varying difficulty of different aspects of stage 2 on sequence of processing stages with a short SOA. Top Panel: increasing difficulty of stage 2B simply adds a constant to RT2 (the same constant would be added at a long SOA, where the two tasks are effectively sequential). Bottom Panel: increasing difficulty of stage 2A does not affect second response time at all, if the SOA were very long, however, this would not be the case.

The bottom panel of Fig 4.4 shows what to expect if we increase the duration of stage 2A by *k* ms. As long as stage 2A still ends before stage 1B, the factor effect onto RT2 will be washed out. In operations research, this is known as "absorption into slack". Following Schweickert (1978) we can call the mental wait state in the diagram "cognitive slack", and the disappearance of factor effects on stage 2B phenomena at short SOAs can then be dubbed "absorption into cognitive slack".

Several predictions follow from this analysis. Overadditive interactions of factor effects and SOA (i.e. larger difficulty effects at shorter SOAs) are not predicted by the mechanism described here. Rather, absorption into slack (underadditivity) is expected for factors affecting early stages of Task 2, while additivity is predicted for factors affecting late stages. In fact, overadditive interactions seem rare in the literature. In the Pashler and Johnston (1989) study, the predictions of underadditivity and additivity were confirmed. In this study, the effect of altering the intensity of Task 2 stimuli, which plainly should affect an early stage of processing, was almost entirely absorbed into slack at short SOAs. An effect of intensity of about 30 ms was reduced to only 5 ms. We also found that the effect of repetition of Task 2 stimulus, which mainly affects the more central stage of S-R mapping (Pashler & Baylis, 1991) was additive with SOA. Underadditive interactions involving manipulations that affect early stimulus processing have also been reported by Johnston, McCann, and Remington (1996) and de Jong (1993).

Additive effects for other manipulations of central decision-making and response-selection operations have been reported by McCann and Johnston (1992), Ruthruff, Miller, and Lachmann (1995), Van Selst and Jolicoeur (1995), McCann and Johnston (1989), and Carrier and Pashler (1996). Dutta and Walker (1995) and McCann and Johnston (1989) found that this additivity persists over at least several sessions of practice, so the bottleneck is not caused by any simple failure to remember the task instructions.

In summary, recent research has provided new and subtle evidence for the existence of a central bottleneck in dual-task performance, at least at low and intermediate levels of practice in "ordinary" choice RT tasks (a distinction that will be clarified later under the heading "Exceptions to the PRP effect"). So far, there does not appear to be any solid body of data that poses substantial problems for bottleneck theory. Obviously, new data might change this situation and force revisions of the model; it seems implausible, however, that any more refined model will have no resemblance to what we have been describing so far.

Locus of PRP Bottleneck

We have seen that there is considerable evidence that a central bottleneck is a principle cause of PRP interference. "Central" is a vague notion, however,

and it would certainly be desirable if the bottleneck could be pinpointed with more precision than that. The main consideration that led Welford (1952) to hypothesize a central bottleneck years ago was the fact that PRP interference arises with pairs of tasks that do not use the same input modality (e.g. one task may be visual and the other auditory). Since then, it has been found with various combinations of different output modalities, too (e.g. a manual response and a vocal response). This suggests, but does not prove, that the interference originates in processes that are neither perceptual nor motoric in character.

The methodology described in the preceding section suggests a way to pinpoint the locus of interference with more precision: by manipulating the time taken for a particular mental operation, one can determine whether this operation is subject to the central bottleneck. When applied to the second task in the PRP design, if the variable has effects that combine additively with SOA, then it must affect a stage at or after the point where the bottleneck begins (or to be more precise, there can be no effective bottleneck[3] after that stage). It seems natural to start with choice reaction time tasks, hut the method can be perfectly well applied to more complex and interesting cognitive tasks that require stages and operations not present in choice reaction time.

An early attempt to use this logic was carried out by Keele (1973), who compared RT2 for simple and choice RT tasks (as Donders noted over a century ago, RTs are slower for choice than for simple RT). Karlin and Kestenbaum (1968) had found that difference between the two was smaller when the task was placed second in the PRP paradigm than when the tasks were performed in isolation. Keele assumed that the RT difference between choice and simple RT reflected "decision" or "memory retrieval", and inferred that these operations were not delayed by the first task. Hence, he concluded that the PRP effect must originate from a delay in initiating or producing responses (rather than choosing them, as suggested earlier). This reasoning can be questioned. First, there is little reason to assume that simple and choice RT differ only in the presence of any single stage or stages (see Goodrich et al., 1990; Pashler, 1994). Second, the issue may now be moot in any case; Van Selst (1996) made repeated attempts to replicate Karlin and Kestenbaum's finding of underadditivity, but instead found additivity.

A number of other variables in choice RT have been investigated more recently using this design. As described earlier, Pashler and Johnston (1989) found that stimulus repetition effects in a choice RT task were additive. As repetition is likely to affect response selection (Pashler & Baylis, 1991), this

[3]By "no effective bottleneck", we mean no bottleneck that is producing slowing within the experiment itself.

suggests that the bottleneck is at or prior to that operation (translation, in Welford's terminology). McCann and Johnston (1992) found additive results using another factor likely to affect the duration of the response selection stage—the "naturalness" of the S-R mapping rule. For instance, in one of their experiments, three sizes of objects were either mapped onto the fingers of one hand in a consistent order (e.g. 1-2-3) or an arbitrary order (e.g. 3-2-1). In another experiment, naturalness of the S-R mapping was manipulated by having subjects respond with a left or right key press to either an arrow (natural mapping) or the letters M and T (arbitrary mapping). Both experiments found that the natural mapping was faster than the unnatural mapping by the same amount at short SOAs (where the PRP delayed RT2) as at long SOAs where the bottleneck did not occur.

So far, then, we can conclude that there is, in choice tasks at least, no bottleneck *after* the S-R mapping stage. This, of course, is fully compatible with the possibility of a bottleneck located at the stage of S-R mapping itself and/or a bottleneck at some prior stages. The fact that varying the difficulty of central processing in *the first* task causes a delay in R2 (Pashler, 1984; Smith, 1969) rejects the idea that the bottleneck is completely prior to response selection. We are left, therefore, with two alternatives: the bottleneck is only in response selection, or it is in response selection plus certain earlier operations—presumably those concerned with stimulus analysis.

Recent evidence suggests that when stimulus analysis is made sufficiently difficult (in certain ways, at least), it too can become subject to the central bottleneck. Johnston and McCann (submitted) used a task requiring an "analog" (i.e. nondiscrete) perceptual judgment as Task 2. In one of their experiments, this task involved pressing one of two keys depending on whether a cross was to the left or right of the center of a circle. Here, the S-R mapping process (i.e. response selection) can be assumed to take as input the left-vs-right classification. Johnston and McCann varied the difficulty of deriving this code by putting the cross either slightly or quite substantially off center. At long SOAs, this variable produced an 88 ms increase in RT2. At short SOAs, this was reduced to 64 ms. Hence, less than one third of the difficulty effect was absorbed into slack. The fact that absorption into slack was far from complete suggests that a considerable amount of input classification was held up by the bottleneck.

In follow-up experiments, Johnston and McCann used a different Task 2: judging whether a rectangle was fat or thin. Four progressively wider stimuli were used—"very thin", "somewhat thin", "somewhat fat", and "very fat". Subjects had to classify the first two as thin and the other two as fat. Responses were substantially faster for the two extreme stimuli compared to the intermediate cases; this difference was the difficulty manipulation of interest. In these experiments, the Task-2 difficulty effect was very close to additive with SOA, implying postponement of the corresponding processing

stage(s). Thus, for this task at least, the bottleneck seems to include operations in Task 2 that occur before S-R translation. It has also been found that when perceptual processing on Task 2 includes complex mediating operations such as mental rotation, these operations are (usually or largely) unable to proceed during the bottleneck (Ruthruff et al., 1995).

In another set of studies, McCann and Johnston (1989 and submitted) examined variables that prolong letter identification. While intensity effects are fully absorbed, as described earlier, these effects might be affecting stages of visual analysis prior to actual identification. McCann and Johnston (submitted) squeezed letters to make them either very squat or very narrow, while keeping the stroke widths and contrast constant. In another experiment, the tilt of component strokes (for instance the diagonal segments in the letter "A") were rotated inward so the character looked a bit like a teepee. The task required subjects to identify the letters. At long SOAs, both experiments showed about a 30 ms slowing of RT2 for the distorted forms. At short SOAs, however, there was complete absorption into slack, i.e. RTs for distorted and undistorted were no different. These results show that some stimulus processing beyond primitive visual feature extraction can occur on Task 2 while critical stages of Task 1 are executed.

We are led to conclude, therefore, that the bottleneck ordinarily encompasses response selection in choice RT tasks and, when present, certain other perceptual operations as well, such as analog comparisons and mental rotation. Some caution is in order, however, as only a fairly modest number of results are available using this method. Like a paleontologist with only a few skulls at hand, we should be cautious about jumping to broad conclusions on the basis of a few studies. The locus-of-slack method is based on assumptions that cannot presently be tested wholly independently of the method. Nonetheless, the method so far has produced consistent and sensible patterns of results, namely underadditive interactions—indicating absorption into slack—for factors one would independently have judged to be perceptual (e.g. visual degradation, letter identification), and additive effects for factors independently believed to affect S-R mapping. This pattern follows naturally from the central bottleneck model, and it is hard to see which other accounts would predict it.

PRP AND OTHER ASPECTS OF ATTENTION

What is the relationship of the central bottleneck revealed in the experiments just described and the very general notion of Attention? As noted at the beginning of this chapter, attention is a diffuse concept. Perhaps the most prominent idea associated with the term is perceptual selectivity: our ability to choose one object from among many for "awareness", memory, and the control of action. In trying to relate the central bottleneck to

attention it seems natural to start by asking whether the machinery that controls perceptual selection plays some role in our inability to choose different responses at the same time. For example, if shifts in attention were carried out by the same central machinery as response selection and other centrally demanding operations, attentional shifts should represent a bottleneck along with response selection and other mental operations.

Several recent sets of studies offer converging ways of addressing this question. One set of experiments used a hybrid methodology, with a speeded first task and an unspeeded second task (Pashler, 1991). The first task required a rapid choice response to a tone, while the second required the observer to shift attention within the visual field based on the nature of a cue presented by the experimenter. In this second task, the display of letters contained an arrow or bar indicating a single item that the observer should attempt to remember and later, at his or her leisure, report. The entire display was followed by a mask. To perform this second task beyond a minimum level[4] one must shift attention to the cued location and store the appropriate item in short-term memory; because the mask appears so quickly, there is barely enough time to complete this on most trials. As in a PRP experiment, the key manipulation was SOA; at very short SOAs, the attention shift would be called for just as the planning of a response to the tone was under way, whereas at long SOAs, it would not have to begin until the first task had been completed (at least on most trials). The results showed, however, that the accuracy of letter report was unaffected by SOA. This implies, then, that response selection and attention shifting occurred in parallel, and thus that the central bottleneck does not encompass shifting visual attention.

Converging evidence for this conclusion comes from recent work by Johnston, McCann, and Remington (1996). Recall that the PRP results described earlier argue that letter identification is not normally subject to the central bottleneck, i.e. identification of letters in Task 2 can proceed without waiting for central processing in Task 1 to be completed (McCann & Johnston, 1989; Johnston et al., 1996). Johnston et al. carried out several new experiments using essentially the same logic to see if letter identification is held up by spatial attention delays. Their results indicated that it is. These results suggest a model in which letter identification requires spatial attention (and therefore occurs *after* spatial attention in tasks where spatial attention is initially directed away from the letters) but does not require whatever limited machinery underlies the central bottleneck. The fact that letter identification would wait for spatial attention but does not have to wait for this central machinery strongly suggests that the two reflect different underlying systems or mechanisms; Johnston, McCann and

[4] The minimum level is actually higher than chance, because people may unselectively store as many items as they can (Sperling, 1960).

Remington (1995) dubbed these "input attention" and "central attention". This conclusion fits nicely with the results of the hybrid speeded/unspeeded task results described in the previous paragraph. Using the terminology of Johnston et al., those results could be described as showing that input attention can be shifted while central attention is occupied with a different task. Both results imply the same fundamental point: that input attention cannot *be* central attention.

Does PRP Slowing Reflect Only a Single Bottleneck?

So far, we have talked as if the PRP effect stems from a single bottleneck (i.e. there is one set of operations constrained so that no operation in that set can overlap with any other in the set). It would obviously be unwise to assume this, however; as with a complex computer system, the brain might have various processors each individually capable of handling only one input at a time. If so, there could be various processor conflicts resulting in various bottlenecks. If the dual-task methodologies described here are applied to a wide variety of tasks involving complex spatial, linguistic, and other cognitive functions, as it seems likely they will be, it is quite possible that a number of bottlenecks will be discovered.

At present, however, a single bottleneck seems sufficient to account for the response delays observed in "standard" PRP designs involving pairs of choice RT tasks. In fact, results from these paradigms are difficult to square with the existence of multiple bottlenecks. The reason is that, as noted earlier, any Task-2 factor that slows a stage prior to the *last* bottleneck will show absorption into slack. Given the robust patterns of additivity observed in many studies, there cannot be "slack" *beyond* the stage manipulated. Nonetheless, there is reason to believe that the processes involved in producing distinct manual responses have the potential to conflict under certain circumstances, introducing a second, independent source of interference. It seems likely, however, that this potential is latent rather than actual in most experiments. To see why, consider some results by Pashler and Christian (unpublished), which involved a PRP task. The duration of the Task-1 response was manipulated: some stimuli required just a single keypress, while others required a series of several keypresses. Naturally, it took more time to finish the series than it took to perform a single keypress. When R2 was a vocal response, however, RT2 was barely affected by this manipulation; on trials where R1 was a sequence, R2 was often emitted while R1 was still under way. Thus, executing the manual responses in Task 1 did not hold up Task 2 (consistent with the idea that the central bottleneck encompasses response selection not response production). On the other hand, when R2 was a (right-hand) *manual* response, R2 was generally held up until the left-hand response sequence was completed. Thus, it seems that producing a

stream of manual responses can potentially engender a second bottleneck in addition to the central bottleneck. However, in the typical PRP experiment with two discrete keypress responses, the first keypress is completed so quickly that there is no conflict between R1 and R2, making this limitation moot.[5] Naturally, daily life may include many situations where this conflict is not moot and does affect performance.

Why a Bottleneck?

So far we have talked about the existence of bottlenecks and their locus, but little has been said about *why* the human mind should be limited in this fashion. We can do little more than speculate here, but some of the possibilities are quite interesting. One interpretation of a bottleneck—so obvious that it is sometimes assumed to be equivalent to the concept of a bottleneck—is that the brain contains only a single piece of machinery capable of performing the critical operations that generate the bottleneck (response selection, retrieval, decision making, etc.). This is a parsimonious interpretation, but not the only possible one, and it may not even be the most plausible. For many decades it has been known that the brain relies on massive parallel processing at the level of individual neurons. There is also good reason to believe that different cortical areas are heavily specialized for carrying out different cognitive functions. However, different cortical areas are generally interconnected, so even if different mental operations are carried out in different neural modules, inhibitory interactions between modules might prevent parallel processing. That is, there might be a "lockout" operation whereby activity in one area suppresses processing in other areas.

Why would such a lockout occur? One possibility is that it might be a functional adaptation designed to prevent crosstalk or confusion. As noted in the introduction, some researchers have argued that crosstalk is *the* essential source of dual-task interference. In the PRP design, at least, this does not fit the facts: pairs of tasks show extreme interference (queuing of central processing) even when they use different input and output modalities, and involve completely different "domains" of cognition (e.g. spatial vs verbal vs lexical information processing; see Pashler, 1994). It is possible, however, that mutual inhibition might occur even with completely dissimilar tasks precisely because the possibility of crosstalk cannot always be excluded; that is, processing lockout might occur whether or not crosstalk actually arises simply because it *might* arise.

Another possibility is that attentional limitations in the control of action may exist to prevent the organism from producing incompatible motor

[5] One exception may be when the tasks are simple rather than choice RT (see De Jong, 1993).

actions, presumably with harmful results. Allport (1989, p. 649), for example, hypothesized the need to maintain "coherent and univocal control of action". One problem with this perfectly reasonable-sounding idea, at least as an explanation for the central bottleneck, is that in many dual-task experiments, subjects *can* in fact execute distinct motor plans at the same time. Indeed, casual observation confirms that people often talk while picking things up, for example. What seems to be blocked is simultaneous planning, rather than simultaneous execution, of motor responses. It is not clear how this would prevent incoherent or conflicting actions. Furthermore, it is not clear that there really is any general process that prevents incoherent or conflicting actions. While motor responses are being produced they are susceptible to both modification and inhibition, although naturally modifications are not instantaneous (Logan & Burkell, 1986). Furthermore, there seems to be no general constraint that prevents people from "willing" and initiating two physically incompatible actions in rapid succession; when this happens, the result is in some ways a compromise of the two (de Jong, 1995). Late modifications of ongoing behavior may, from time to time, cause what we would judge to be action errors, but these need not necessarily have disastrous consequences. Baars and Motley (1976) have argued persuasively that Spoonerisms and other speech errors often occur when speech plans are changed after the plans have already been partly carried out.

EXCEPTIONS TO THE PRP EFFECT?

If there were no fundamental constraint preventing central stages of multiple tasks from being carried out simultaneously, one might expect that exceptions to PRP interference would be encountered frequently. But in tact, amidst the hundreds of studies in this area, only a handful of exceptions have been noted. These examples appear to be cases where two conditions are met: (1) modality conflicts in perceiving and responding are avoided, and (2) one or both tasks are unusually "easy" in some intuitive sense. (In fact, these two conditions are probably necessary but not sufficient to avoid PRP interference.) These exceptions have generally been interpreted as indicating that certain specific neural pathways are capable of bypassing the central bottleneck.

Greenwald (1972) and Greenwald & Shulman (1973) reported finding virtually no PRP interference when a special relationship between stimuli and responses was present in both tasks. For example, in Greenwald and Shulman (1973), Task 1 required moving a switch right to a right arrow and left to a left arrow, while Task 2 required subjects to say "A" to the spoken signal "A", and "B" to the spoken signal "B". Greenwald and Shulman suggested that these tasks produced little PRP interference because they did

not require the normal process of mapping arbitrary stimuli onto responses. They argued that in these experiments, the stimuli generated a mental code that was already in the right format to select the response. It is certainly not implausible that there is some direct relation between the stimulus and response codes for "right" and "left", and it has long been suspected that listening and speaking speech codes for the same word are closely connected. McLeod and Posner (1984) found similar results when one task involved immediate verbal repetition (shadowing) and argued that there may be a small number of "privileged loops" that bypass the main path for stimulus-response mapping.

A few other exceptions have been reported more recently. Pashler, Carrier, and Hoffman (1993) found that PRP essentially disappeared when Task 2 required shifting eye position to the location of a stimulus. Johnston and Delgado (1993) found virtually no PRP when Task 2 was a simple tracking task in which subjects controlled the position of a circle and tried to keep it over a moving stimulus cross. The dynamics required leftward or rightward movements of a joystick in response to leftward and rightward movements of the stimulus cross. Note that both the eye movement task and the tracking task involved extremely "natural" stimulus-response mapping rules.

Johnston and Delgado (1993) found that PRP interference could be avoided when tracking served as Task 2 even though Task 1 required an arbitrary mapping of spoken stimulus words to spoken response words. The authors speculated that what is critical for avoiding a PRP is only that the second task be able to bypass the bottleneck channel; if it can, it should not matter whether Task 1 occupies that channel or not. In a more speculative vein, Johnston and Delgado suggested that there might be two requirements for avoiding PRP interference. The first is that subjects be able to conceive of Task 2 responses as "pre-authorized" prior to any particular movement. In the tracking task, no new "respond now" command is required for each response. Second, in line with previous theorizing, there must be a natural mapping of stimulus and response codes so that lower-level systems can carry out the pre-authorized responses without central mechanisms. In the case of tracking, what is required is that the brain use closely related spatial coordinate systems for the location of stimuli and the location of the action-consequences of responses. If either of these conditions is not met, higher-level systems will be needed either to authorize responding or to accomplish the S-R mapping, and a PRP will be produced.

WHERE IS THE BOTTLENECK ANATOMICALLY?

Although purely behavioral methods have taken us a considerable distance, it would obviously be desirable if our partial functional understanding of dual-task performance limitations could be tied to underlying brain structures and processes. Eventually one might hope to be able to record neural events indicative of bottlenecks and capacity sharing. This goal is as yet unrealized, but an initial step was taken in a recent study that measured event-related potentials during a PRP experiment (Osman & Moore, 1993). These authors found that the lateralized readiness potential over motor cortex associated with R2 was delayed to approximately the same extent as the RT itself. The result suggests that the lateralized readiness potential occurs at the completion of response selection, a view that seems to fit a number of other findings.

Other clues about the neural locus of PRP interference have emerged from recent studies of an unusual class of neurological patients—so-called "split-brain" patients in whom the connections between the cortical hemispheres have been surgically severed. If the central bottleneck described earlier has a cortical locus, split-brain patients should have two separate response-selection devices and exhibit no PRP effect whenever the two tasks are confined to separate hemispheres. However, using lateralized stimuli and responses, Pashler et al. (1994) observed a more or less normal PRP effect in four split-brain patients. They suggested that the structures causing the PRP bottleneck might therefore have a sub-cortical source, as connections at these brain levels remain intact in split-brain patients.

Examining one of the four split-brain patients studied by Pashler et al. (patient JW), Ivry, Franz, Kingstone, and Johnston (in press) confirmed the occurrence of PRP interference even when different input and response modalities were used for the two tasks. However, they have also obtained findings suggesting that the bottleneck may have a different locus for patient JW than it has for normals. In one experiment, Ivry et al. tested for the effects of inconsistent mapping rules in each task, using a method first employed by Duncan (1979). Both tasks required subjects to press an upper or a lower key in response to an upper or lower light. Sometimes, however, the S-R mapping for a task was the natural "compatible" mapping (upper light to upper button, lower light to lower button), whereas sometimes it was an unnatural, "compatible" mapping (upper light to lower button, lower light to upper button). Not surprisingly, subjects took longer to make a response when they used the compatible mapping than when they used the incompatible mapping. In a dual-task experiment with normal subjects, Duncan found that when the compatible mapping must be used on one task and the incompatible mapping on the other, responses in both tasks are slowed. In a PRP design, one would expect that having to switch from one mapping rule to another would slow Task 2 responses, as indeed it does. Interestingly, the inconsistency also slows responses on the *first* task; evidently, the alternation between rules impairs preparation of both tasks.

With patient JW, Ivry et al. used two conditions from Duncan's design. In one condition both Task 1 and Task 2 used compatible mappings (hence a

consistent relationship between the mappings). In the other condition, Task 1 used a compatible mapping and Task 2 used an incompatible mapping. Normal subjects showed results like Duncan's. Although the only difference between the two conditions was in the mapping rule used in Task 2, both RT1 and RT2 were substantially lengthened. However, for patient JW changing the mapping rule on Task 2 to an incompatible mapping substantially lengthened RT2 but had almost no effect on RT1. Thus it appears that for this patient, the hemisphere doing the compatible mapping was not affected by the fact that the other hemisphere was using an inconsistent mapping rule. There was one further difference between normals and patient JW. The variation in mapping rule on Task 2 constitutes a Task 2 difficulty variable so we can apply the locus of slack analysis. As this manipulation presumably affects the duration of the S-R mapping stage we should expect to find no absorption into slack, as in several experiments described earlier. This is essentially what Ivry et al. found for the normal subjects. For patient JW, however, the effect of the change in Task 2 mapping on RT2 was about 200 ms at the longest SOA but only about 40 ms at the shortest SOA. Thus almost all of the effect of the task 2 S-R mapping variable on RT2 was absorbed into slack. This indicates that for this patient—but not for normals—the bottleneck locus is after the S-R mapping stage.

Pashler et al. had inferred that as patient JW and normals both show a PRP bottleneck, this bottleneck is probably in an anatomical structure that is intact in JW; hence that the normal bottleneck is subcortical in origin. However, the new results challenge this conclusion, by suggesting that the bottleneck in patient JW may arise at a different processing stage than it does in normals. Therefore we cannot necessarily assume that the anatomical locus of the bottleneck in patient JW corresponds to the anatomical locus of the bottleneck in normals. There are at least two possible interpretations of these results.

The first is that patient JW's data reveal a bottleneck that exists in normals, but is normally "latent", in very much the same sense that manual response production bottleneck was argued to be latent in the typical PRP experiment with dual manual responses.[6] That is, as patient JW has no bottleneck at the stage of response selection (a conclusion suggested by the virtual elimination of the inconsistency effect in this patient), a bottleneck at a later stage (perhaps some brief stage of initiating responses) emerges—a bottleneck that is normally concealed in normals by queuing of the earlier response selection stage.

[6] To be slightly more formal: even if stage J in Task 1 and stage J in Task 2 cannot be performed simultaneously, this may be only a latent conflict if, due to the timing of preceding stages, and due to possible bottlenecks arising in these earlier stages, it will never happen that the input to stage J in Task 2 will be ready before stage J in Task 1 has been completed.

A second possibility is that the separation of patient JW's cortical hemispheres may have removed all central bottlenecks, removing any internal obstacles to parallel performance on both tasks. However, as JW was studied years after his surgery, he may have acquired novel strategies to prevent the two sides of his brain from acting incoherently. For instance, what would happen if one side of his brain decided it wanted to continue moving a spoonful of soup to his mouth to eat, while the other side decided to command the other hand to push away from the table so he could get up and do something else entirely? Conceivably, patient JW has had to learn some strategy for letting his hemispheres take turns at controlling his overt actions, a strategy that carries over to his performance in the PRP paradigm. Ironically, then, the suggestion that central bottlenecks are not structural hut merely strategic may have some validity, but only for split-brain patients and not for intact individuals.

Although the attempt to characterize the source of PRP interference in functional terms has been going on for about half a century, attempts to uncover the physiological/anatomical locus of central interference has barely begun. With recent advances in brain imaging technologies, we may be able to look forward to new insights on this question.

ATTENTIONAL LIMITS IN MEMORY

Many theorists have assumed that attentional capacity limitations can be equated in one fashion or another with short-term memory (STM). Although some writers have questioned the empirical validity of the concept of short-term memory, there is a formidable body of evidence for the idea that information can be stored in a transient form that is distinct from permanent memory (see Pashler & Carrier, 1996, for a review). This evidence argues for a multiplicity of short-term memory systems rather than a single system, however. There is also evidence that short-term storage usually reflects brain systems specialized for carrying out specific cognitive functions aside from short-term memory (e.g. language comprehension, motor control, visual perception). For this reason, we use the term "STM" here to refer to various different systems in which information can be held transiently—systems that may well have other functions besides memory.

It has long been supposed that there is a close and important connection between "attention" and short-term retention. For one thing, people have selective control over what information they hold onto for immediate report. Some illustrations of this are found in the classic partial-report experiments involving audition (Darwin, Turvey & Crowder, 1972) and vision (Sperling, 1960). Here, subjects were able to transfer items into short-term memory on the basis of some cued attribute like the position or color of a letter.

What sorts of attentional limitations is STM subject to? One way to address this question is to see whether information can be transferred into STM while a person carries out a centrally demanding operation in an unrelated task—e.g. planning an action of some kind. Is the transfer of information into STM subject to the same central bottleneck responsible for the PRP effect described in the preceding section?[7] A number of investigators have presented lists of words while people performed a concurrent task, and later tested memory for the word using the "free recall" task (repeating the items back in any order). Murdock (1965), for example, had subjects listen to a spoken list while rapidly sorting cards, and then attempt free recall. The so-called "recency effect" (memory for the last-presented items) was quite intact, suggesting that the sorting task did not prevent words from being stored in STM. Anderson and Craik (1974) made similar observations using a list of spoken words presented while subjects performed a concurrent visual/manual choice reaction-time task.

Other evidence also argues that storage in visual short-term memory is relatively free of central capacity demands. One experiment that explored this issue required subjects to make a rapid choice response to a tone; at some point during or after the choice task, a pattern of black and white squares was flashed briefly, followed immediately by a mask. Subjects were able to maintain good performance regardless of the temporal overlap of the two tasks, suggesting that information about the patterns was stored in visual STM while the response to the tone was being planned (Pashler, 1993b).

What about simply holding onto information already in STM? Passively retaining a memory load slows responses in concurrent speeded tasks to some extent (Logan, 1978). However, it does not severely impair performance in difficult reasoning and comprehension tasks (Baddeley & Hitch, 1977). It is an odd fact that researchers have very frequently assumed that STM storage drains "general processing resources". The results just described would seem to test this assumption and reject it. On the other hand, beginning to rehearse information that has just been stored in STM (which intuitively feels more like carrying out an action rather than merely maintaining a state) does seem to produce substantial interference that lasts for a short time (Naveh-Benjamin & Jonides, 1984).

What about long-term memory (LTM)? As with short-term memory, when someone deliberately ignores a stimulus there is often little trace of that stimulus evident in later tests of long-term memory (Moray, 1959; Rock & Guttman, 1981). However, unlike with STM, concurrent tasks that

[7] The alert reader will notice that evidence already described suggests that there could not be any complete interference of this sort, in the experiments combining speeded responses with attention shifts.

impose central processing demands clearly reduce the flow of information into LTM. Evidence arguing for this conclusion comes from experiments described earlier in which people listened to words while carrying out concurrent sorting tasks. Although recency (reflecting predominantly STM) was unaffected, recall of items from earlier parts of the list (reflecting predominantly LTM) was much reduced (Anderson & Craik, 1974; Murdock, 1965). Further evidence comes from recent studies in which people carried out speeded choice tasks while items to be remembered were presented (Carrier & Pashler, unpublished data). When the concurrent tasks involved speeded responses to tones, and there was a short interval between the response to a tone and the occurrence of the next tone, subsequent memory for material read during the tone task suffered substantially. This was true whether the material consisted of word lists or sequences of faces, and whether memory was measured with recall or recognition. This result reinforces the earlier observation that profound dual-task interference may occur without the materials involved in the two tasks being discernibly similar. This suggests competition for a relatively general mechanism or some process of mutual inhibition.

The fact that the flow of information into long-term memory is impaired by concurrent central demands would suggest that memory storage is subject to the very same central bottleneck as the PRP. However, measures of accuracy of memory storage, unlike reaction time, do not tell us whether the concurrent task completely prevents the memory storage, or merely slows it. Empirically, dual-task manipulations generally reduce memory performance without bringing it down to chance levels. Before leaving this topic, it should be mentioned that some reports in the literature have concluded that secondary tasks do not reduce memory storage. Usually, these studies have involved concurrent tasks that require only intermittent central processing. For example, Tun, Wingfield, and Stine 1991) found that a concurrent choice reaction-time task did not reduce later recall of spoken prose passages. However, the concurrent task involved responding to letters that were presented only once every three to seven seconds, a task that would be expected to occupy central processing machinery for only a tiny fraction of the total time.

So far, we have talked about storing information in memory; we turn now to *retrieving* information that has already been stored in long-term memory. Experiments requiring people to carry out memory retrieval together with a concurrent task have led to conflicting conclusions. On the one hand, Park, Smith, Dudley, and Lafronza (1989) found that an auditory/manual concurrent task impaired concurrent (verbal) free recall. On the other, Baddeley et al. (1984) combined a sorting task with memory retrieval, and found that the difficulty of the sorting had little effect on the success of retrieval, although it did increase response latencies. Baddeley et al.

concluded that interference affected the production of responses but not the actual memory retrieval process.

More fine-grained analyses are necessary to discriminate between interference in production and interference in retrieval. As described earlier, the PRP method is particularly well suited for this. Carrier and Pashler (1996) combined a manual response to a tone (Task 1) with retrieval of a paired associate in response to a visually presented cue word (Task 2) in a PRP design. There was a PRP effect (slower responses in the paired-associate task at shorter SOAs). The duration of the memory retrieval was manipulated in various ways. These manipulations slowed second-task RTs, and the slowing was roughly additive with SOA (Carrier & Pashler, 1996). Following the logic described earlier, this implies that memory retrieval, but not response production or any other stage in the second task after memory retrieval, was delayed when the first task was being performed. This in turn suggests that the central bottleneck described earlier encompasses memory retrieval. One is led to suspect, then, that the inability to select two responses at the same time (response selection bottleneck) is just a special case of a general constraint on retrieving associations in memory, not a limitation particularly tied to motor programming or action.

CONCLUDING COMMENTS

One general theme of this chapter is that examining dual-task performance in detail reveals that our cognitive machinery is subject to more severe limitations than we might have suspected from casual observation. Although perceptual machinery seems capable of identifying more than a single object at a time, it is subject to capacity limits that become evident when the stimulus load is increased beyond a fairly modest level. In the realm of memory retrieval and action planning, a different and more central form of limitation seems to arise. The evidence presently available suggests that overlap in the central operations of different tasks simply does not occur except for a few special cases of extremely compatible stimulus-response mappings. Sufficient practice may get around this limitation, but this has not yet been demonstrated; several thousand trials of practice in rather simple choice reaction-time tasks seem insufficient.

It should be noted that the idea of obligatory serial central processing is quite consistent with a great deal of parallel processing, for several reasons. First, central processing in one task can clearly overlap with both perceptual analysis and production of motor responses in another task (as Fig. 4.2 makes plain). Second, planning of a response to a given stimulus seems to overlap with continuing perceptual analysis of that stimulus (Levy & Pashler, 1995). Thus, at the very moment we are planning an action based on preliminary perceptual conclusions about some objects, these conclusions may be refined and even overturned. This implies that the boxes in stage diagrams of the sort shown in Fig. 4.2 should be seen as depicting a stream of processing that results in a given response, not the totality of the processing of a given stimulus that the nervous system may carry out. Third, at least in the case of dissimilar motor responses (e.g. vocal and manual responses), two independent streams of outputs can often be produced. This is demonstrated in the continuous-task experiments described at the beginning of the chapter, and confirmed in the PRP experiments that combined sequences of responses in one task and punctate responses in a second task (Pashler & Christian, unpublished). Fourth, even though the use of different inputs to select different responses to each input requires sequential processing, when several inputs select a *single* response this lookup can be carried out in a single mental operation. This is demonstrated in the so-called coactivation effect observed by Miller (1982), and is also seen in people's ready ability to solve crossword puzzles using completely unrelated cues to "home in on" a target in memory.[8]

Therefore, the central bottleneck argued for in this chapter does not conflict with casual observations that people often read a newspaper while riding an exercise bicycle, for example, or move a cup of coffee away from their lips while speaking. Nor does it conflict with the idea that different brain areas—some specialized more for stimulus-related processing, some more for response-related processing—usually work continuously and concurrently. What these results do suggest, however, is that in certain important respects our mind may nonetheless work a bit like a digital computer with switching and buffering capabilities, and that fine-grained measurements of performance in dual-task situations can reveal many non-obvious facts about the timing of underlying mental/neural events.

REFERENCES

Allport, A. (1989). Visual attention. In M.I. Posner (Ed.), *Foundations of cognitive science* (pp. 631-682). Cambridge, MA: MIT Press.

Allport, D.A., Antonis, B., & Reynolds, P. (1972). On the division of attention: A disproof of the single-channel hypothesis. *Quarterly Journal of Experimental Psychology, 24,* 225-235.

Anderson, C.M.B., & Cralk, F.I.M. (1974). The effect of a concurrent task on recall from primary memory. *Journal of Verbal Learning and Verbal Behavior, 13,* 107-113.

Baars, B.J., & Motley, M.T. (1976). Spoonerisms as sequencer conflicts: Evidence from artificially elicited errors. *American Journal of Psychology, 89,* 467-484.

Baddeley, A., Lewis, V., Eldridge, M., & Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General, 113,* 518-540.

Baddeley, A.D. (1986). *Working memory.* Oxford: Oxford University Press.

Baddeley, A.D., & Hitch, G. (1977). Recency reexamined. In S. Dornic (Ed.), *Attention & performance VI* (pp. 647-667). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

[8]For a discussion of the implications of this for cognitive architecture, see Pashler (1993a).

Bertelson, P. (1966). Central intermittency twenty years later. *Quarterly Journal of Experimental Psychology, 18,* 153-163.

Biederman, I., Teitelbaum, R.C., & Mezzanotte, R.J. (1983). Scene perception: A failure to find a benefit from prior expectancy or familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 411-429.

Carrier, M., & Pashler, H. (1996). The attention demands of memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1339-1348.

Darwin, C.J., Turvey, M.T., & Crowder, R.G. (1972). An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. *Cognitive Psychology, 3,* 255-267.

de Jong, R. (1993). Multiple bottlenecks in overlapping task performance. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 965-980.

de Jong, R. (1995). Perception-action coupling and S-R compatibility. *Acta Psychologica, 90,* 287-299.

Duncan, J. (1979). Divided attention: The whole is more than the sum of its parts. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 216-228.

Duncan, J. (1980a). The demonstration of capacity limitation. *Cognitive Psychology, 12,* 75-96.

Duncan, J. (1980b). The locus of interference in the perception of simultaneous stimuli. *Psychological Review, 87,* 272-300.

Duncan, J. (1987). Attention and reading: Wholes and parts in shape recognition—A tutorial review. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 36-61). Hove, UK: Lawrence Erlbaum Associates Ltd.

Duncan, J. (1993). Similarity between concurrent visual discriminations: Dimensions and objects. *Perception and Psychophysics, 54,* 425-430.

Duncan, J., Ward, R., & Shapiro, K. (1994). Direct measurement of attentional dwell time in human vision. *Nature, 369,* 313-315.

Dutta, A., & Walker, B.N. (1995, November). *Persistence of the PRP effect: Evaluating the response-selection bottleneck.* Paper presented at the 36th annual meeting of the Psychonomic Society, Los Angeles, California.

Eriksen, C.W., & Hoffman, J.E. (1973). The extent of processing of noise elements during selective encoding from visual displays. *Perception and Psychophysics,* 14, 155-160.

Estes, W.K., & Taylor, H.A. (1964). A detection method and probabilistic models for assessing information processing from brief visual displays. *Proceedings of the National Academy of Sciences, 52,* 446-454.

Estes, W.K., & Taylor, H.A. (1965). Visual detection in relation to display size and redundancy of critical elements. *Perception and Psychophysics,* 1, 9-15.

Goodrich, S., Henderson, L., Allchin, N., & Jeevaratnam, A. (1990). On the peculiarity of simple reaction time. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 42A,* 763-775.

Gottsdanker, R. (1980). The ubiquitous role of preparation. In G.E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 355-371). Amsterdam: North-Holland Press.

Gottsdanker, R., & Way, T.C. (1966). Varied and constant intersignal intervals in psychological refractoriness. *Journal of Experimental Psychology, 72,* 792-804.

Greenwald, A.G. (1972). On doing two things at once: Time-sharing as a function of ideomotor compatibility. *Journal of Experimental Psychology, 100,* 52-57.

Greenwald, A., & Shulman, H. (1973). On doing two things at once: II. Elimination of the psychological refractory period. *Journal of Experimental Psychology, 101,* 70-76.

Ivry, R.B., Franz, E., Kingston, A., & Johnston, J.C. (in press). The PRP effect in a split-brain patient: Response uncoupling despite normal interference. *Journal of Experimental Psychology: Human Perception and Performance.*

James, W. (1890/1950). *The principles psychology. Vol. 1.* New York: Dover.

Johnston, J.C. & Delgado, D. (1993, November). *Bypassing the single-channel bottleneck in dual-task performance.* Paper presented at the 34th annual meeting of the Psychonomic Society.

Johnston, J.C. & McCann, R.S. (1997). *On the focus of dual-task interference: Is there a bottleneck at the stimulus classification stage?* Manuscript submitted for publication.

Johnston, J.C., McCann, R.S. & Remington, R.W. (1995). Chronometric evidence for two types of attention. *Psychological Science, 6,* 365-369.

Johnston, J.C., McCann, R., & Remington, R. (1996). Selective attention operates at two processing loci. In A. Kramer & G. Logan (Eds.), *Essays in honor of Charles Eriksen* (pp.439-458). American Psychological Association.

Kahneman, D. (1973). *Attention and effort.* New York: Prentice Hall.

Karlin, L., & Kestenbaum, R. (1968). Effects of number of alternatives on the psychological refractory period. *Quarterly Journal of Experimental Psychology, 20,* 167-178.

Keele, S.W. (1973). *Attention and human performance.* Pacific Palisades, CA: Goodyear.

Kinsbourne, M. (1981). Single channel theory. In D. Holding (Ed.), *Human Skills* (pp. 65-89). Chichester, UK: Wiley.

Kleiss, J.A., & Lane, D.M. (1986). Locus and persistence of capacity limitations in visual information processing. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 200-210.

Koutstaal, W. (1992). Skirting the abyss: A history of experimental explorations of automatic writing in psychology. *Journal of the History of the Behavioral Sciences, 28,* 5-27.

Levy, J., & Pashler, H. (1995). Does perceptual analysis continue during selection and production of a speeded response? *Acta Psychologica, 90,* 245-260.

Logan, G.D. (1978). Attention in character classification tasks: Evidence for the automaticity of component stages. *Journal of Experimental Psychology: General, 107,* 32-63.

Logan, G.D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1015-1036.

Logan, G.D., & Burkell, J. (1986). Dependence and independence in responding to double stimulation: A comparison of stop, change and dual-task paradigms. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 549-563.

McCann, R.S., & Johnston, J.C. (1989, November). *The locus of processing bottlenecks in the overlapping tasks paradigm.* Paper presented at the 1989 meeting of the Psychonomic Society, Atlanta, Georgia.

McCann, R.S., & Johnston, J.C. (1992). Locus of the single-channel bottleneck in dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 471-484.

McLeod, P. (1977). Parallel processing and the psychological refractory period. *Acta Psychologica, 41,* 381-391.

McLeod, P., & Posner, M.I. (1984). Privileged loops from percept to act. In H. Bouma & D.G. Bouwhuis, (Eds.), *Attention and performance X.* Hove, UK: Lawrence Erlbaum Associates Ltd.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14,* 247-279.

Miller, J., & Bonnel, A.M. (1994). Switching or sharing in dual task line length discrimination? *Perception and Psychophysics, 56,* 431-446.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence instructions. *Quarterly Journal of Experimental Psychology, 11,* 56-60.

Murdock, B.B. (1965). Effects of a subsidiary task on short-term memory. *British Journal of Psychology, 56,* 413-419.

Naveh-Benjamin, M., & Jonides, J. (1984). Maintenance rehearsal: A two-component analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 369-385.

Navon, D., & Miller, J. (1987). Role of outcome conflict in dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 435-448.

Osman, A., & Moore, C. (1993). The locus of dual-task interference: Psychological refractory effects on movement-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 1292-1312.

Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision Research, 34,* 1703-1721.

Park, D.C., Smith, A.D., Dudley, W.N., & Lafronza, V.N. (1989). Effects of age and a divided attention task presented during encoding and retrieval on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1185-1191.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology. Human Perception and Performance, 10,* 358-377.

Pashler, H. (1989). Dissociations and dependencies between speed and accuracy: Evidence for a two-component theory of divided attention in simple tasks. *Cognitive Psychology, 21,* 469-514.

Pashler, H. (1991). Shifting visual attention and selecting motor responses: Distinct attentional mechanisms. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 1023-1040.

Pashler, H. (1993a). Doing two tasks at the same time. *American Scientist, 81,* 46-55.

Pashler, H. (1993b). Dual-task interference and elementary mental mechanisms. In D. Meyer & S. Kornblum (Eds.), *Attention and performance XIV (pp.* 245-264). Cambridge, MA: MIT Press.

Pashler, H. (1994). Overlapping mental operations in serial performance with preview *Quarterly Journal of Experimental Psychology, 47,* 161-191.

Pashler, H., & Baylis, G.C. (1991). Procedural learning: II. Intertrial repetition effects in speeded-choice tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 33-48.

Pashler, H., & Carrier, M. (1996). Structures, processes and flow of control. In E.L. Bjork, & R.A. Bjork (Eds.), *Handbook of perception and cognition: Memory* (2nd edn. pp. 3-29). San Diego, CA: Academic Press.

Pashler, H., Carrier, M., & Hoffman, J. (1993). Saccadic eye movements and dual-task interference. *Quarterly Journal of Experimental Psychology, 46A,* 51-82.

Pashler, H., & Christian, C. (1997). *Dual-task interference and motor response production.* Unpublished manuscript.

Pashler, H., & Johnston, J.C. (1989). Interference between temporally overlapping tasks: Chronometric evidence for central postponement with or without response grouping. *Quarterly Journal of Experimental Psychology,* 41A, 19-45.

Pashler, H., Luck, S.J., Hillyard, S.A., Mangun, G.R. & Gazzaniga, M. (1994). Sequential operation of disconnected cerebral hemispheres in split-brain patients. *Neuroreport, 5,* 2381-2384

Pashler, H., & O'Brien, S. (1993). Dual-task interference and the cerebral hemispheres. *Journal of Experimental Psychology: Human Perception and Performance,* 19, 315-330.

Rock, I., & Guttman, D. (1981). The effect of inattention on form perception. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 275-285.

Ruthruff, E., Miller, J., & Lachmann, T. (1995). Does mental rotation require central mechanisms? *Journal of Experimental Psychology: Human Perception and Performance, 21,* 552-570.

Ruthruff, E., & Pashler H. (submitted) Dual-task central bottleneck: Structural or strategic?

Salthouse, T.A., & Saults, J.S. (1987). Multiple spans in transcription typing. *Journal of Applied Psychology, 72,* 187-196.

Schweickert, R. (1978). A critical path generalization of the additive factor method: Analysis of a Stroop task. *Journal of Mathematical Psychology, 18,* 105-139.

Shaffer, L.H. (1975). Multiple attention in continuous verbal tasks. In P.M.A. Rabbitt & S. Dornic (Eds.), *Attention and performance V.* New York: Academic Press.

Shiffrin, R.M., & Gardner, G.T. (1972). Visual processing capacity and attentional control. *Journal of Experimental Psychology, 93,* 78-82.

Smith, M.C. (1967). Theories of the psychological refractory period. *Psychological Bulletin, 67,* 202-213.

Smith, M.C. (1969). The effect of varying information on the psychological refractory period. In W.G. Koster (Ed.), *Acta Psychologica, 30,* 220-231

Spelke, E., Hirst, W., & Neisser, U. (1976). Skills of divided attention. *Cognition, 4,* 215-230.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied* (Whole No. 498, 1-29).

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. In W.G. Koster (Ed.), *Attention and performance II* (pp.276-315). Amsterdam: North Holland.

Telford, C.W. (1931). The refractory phase of voluntary and associative responses. *Journal of Experimental Psychology, 14,* 1-36.

Treisman, A. (1991). Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 652-676.

Treisman, A., & Davies, A. (1973). Dividing attention to ear and eye. In S. Kornblum (Ed.), *Attention and performance IV.* (pp.101-117) New York: Academic Press.

Tun, P.A., Wingfield, A., & Stine, E.A.L. (1991). Speech-processing capacity in young and older adults: A dual-task study. *Psychology and Aging, 6,* 3-9.

Van Selst, M., & Jolicoeur, P. (1995). Can mental rotation occur before the dual-task bottleneck? *Journal of Experimental Psychology: Human Perception and Performance, 20,* 905-921.

Van Selst, M. (in press). Decision and response in dual-task interference. *Cognitive Psychology*

Welford, A.T. (1952). The "psychological refractory period" and the timing of high speed performance: A review and a theory. *British Journal of Psychology, 43,* 2-19.

Welford, A.T. (1967). Single channel operation in the brain. *Acta Psychologica, 27,* 5-22.

Welford, A.T. (1980). The single-channel hypothesis. In A.T. Welford (Ed.), *Reaction time* (pp.215-252). New York: Academic Press.