Running head: PROCESSING MODES IN PRP

# On the Optimality of Serial and Parallel Processing in the Psychological Refractory Period Paradigm: Effects of the Distribution of Stimulus Onset Asynchronies

Jeff Miller

University of Otago

Rolf Ulrich and Bettina Rolke

Universität Tübingen

Address editorial correspondence to Jeff Miller, Department of Psychology, University of Otago, Dunedin, New Zealand, International FAX: 64-3-479-8335, email: miller@psy.otago.ac.nz. *Version of November 13, 2004.*

**Abstract**

Within the context of the psychological refractory period (PRP) paradigm, we developed a general theoretical framework for deciding when it is more efficient to process two tasks in serial and when it is more efficient to process them in parallel. This analysis suggests that a serial mode is more efficient than a parallel mode under a wide variety of conditions and thereby suggests that ubiquitous evidence of serial processing in PRP tasks could result from performance optimization rather than from a structural bottleneck. The analysis further suggests that the experimenter-selected distribution of stimulus onset asynchronies (SOAs) influences the relative efficiency of the serial and parallel modes, with a preponderance of short SOAs favoring a parallel mode. Experiments varying the distribution of SOAs were conducted, and the results suggest that there is a shift from a more serial mode to a more parallel mode as the likelihood of short SOAs increases.

# On the Optimality of Serial and Parallel Processing in the Psychological Refractory Period Paradigm: Effects of the Distribution of Stimulus Onset Asynchronies

The *psychological refractory period* (PRP) paradigm has often been used to study the factors limiting cognitive performance in dual-task situations (e.g., Pashler, 1984; Telford, 1931; Welford, 1952, 1959). In the most typical versions of this paradigm, participants are asked to perform two separate choice reaction time (RT) tasks in each trial. The stimuli for the two tasks—S1 and S2—are presented in rapid succession, and participants are asked to respond to each as quickly as possible.

The PRP paradigm is popular partly because it provides experimenters with fine-grained control over the time interval separating the onsets of S1 and S2, an interval known as the *stimulus onset asynchrony* (SOA). When the SOA is relatively long, participants can simply perform the tasks one after the other, because processing of S1 can finish before S2 is presented. In this case, not surprisingly, the latency of the second response, RT2, is approximately the same as (or only slightly longer than) it would be if that task were performed in isolation. When the SOA is short, however, S1 is still being processed when S2 arrives, and participants must somehow cope with the demands of two simultaneous cognitive tasks. In this case performance generally slows dramatically (for a review see, e.g., Pashler, 1994a). In particular, RT2 increases substantially at short SOAs (Kahneman, 1973), and this increase is generally known as the *PRP effect*. Effects of SOA on RT1 are generally much smaller (e.g., Smith, 1969) and are sometimes essentially absent (e.g., Pashler & Johnston, 1989).

One attractive hypothesis about the cause of the PRP effect is the *response-selection bottleneck model* (Pashler, 1984, 1994b; Welford, 1952, 1959). According to this model, one stage—called the bottleneck—is only capable of processing one task at a time. That is, this stage must process the tasks serially for some structural reason. When the second task needs access to the bottleneck stage while this stage is still busy processing the first task, the second task simply has to wait. Because such waiting time contributes directly to RT2, this model predicts that RT2

should decrease approximately linearly with slope -1 as SOA is increased. Although observed slopes relating RT2 to SOA are often shallower than this (Kahneman, 1973), the observed values are close enough to the predictions for many theorists to conclude that they support the bottleneck model (Pashler, 1994b).

There is still debate about the bottleneck model, however, because other models can also predict that RT2 should increase as SOA decreases, possibly even with a slope of approximately -1. For example, limited-capacity models are often discussed as alternatives to the bottleneck model (e.g., Kahneman, 1973; Navon & Gopher, 1979). The common feature of these models is that processing capacity can be shared between tasks in a graded fashion, with perhaps 70% of processing capacity allocated to one task and 30% to the other. Thus, capacity models are fundamentally different from the bottleneck model in that every stage is capable of processing two tasks in parallel—that is, there is no structural bottleneck.[1] Recent investigations indicate that some versions of these models can predict slopes of approximately -1 and can also accommodate other evidence previously cited as selectively supporting the bottleneck model (e.g., Navon & Miller, 2002; Tombu & Jolicœur, 2003). In addition, several other models allow the possibility of parallel processing, at least under some circumstances (e.g., Logan & Gordon, 2001; Meyer & Kieras, 1997a, 1997b; Navon, 1984). In general, such models seem more capable than bottleneck models of explaining observations that Task 1 responses may be affected by the nature of the response selection required for Task 2 (e.g., Hommel, 1998; Logan & Delheimer, 2001; Logan & Schulkind, 2000).

One reason that it has proved difficult to test experimentally between the bottleneck model and its competitors that allow parallel processing is that the latter models can closely mimic the bottleneck model (e.g., Meyer & Kieras, 1997b; Navon & Miller, 2002; Tombu & Jolicœur, 2003). To our knowledge, virtually all models that allow parallel processing also allow serial processing, so the fact that two tasks *could be* processed in parallel does not imply that they always *would be*.[2] For example, serial processing might be preferred because it is a natural way to bind together the separate sources of information relevant to each task (e.g., Logan & Gordon, 2001) or because it prevents crosstalk between tasks (e.g., Navon & Miller, 1987). The present article emphasizes another possibility: even if parallel processing were possible, people would be unlikely

to use this mode if the serial mode were more efficient. Therefore, theorists should consider the possibility that serial processing leads to better performance than parallel processing before attributing such processing to structural limitations (i.e., a bottleneck).

In this article we focus primarily on the distinction between the bottleneck model, which requires serial processing in a certain stage, and other models that allow parallel processing in all stages. Although a number of studies have been conducted to see whether parallel processing takes place in paradigms designed to encourage it (e.g., Ruthruff, Pashler, & Klaassen, 2001; Tombu & Jolicœur, 2002), none of these studies have presented a theoretical framework that could be used to determine when the serial versus parallel processing modes would be optimal. Instead, in devising paradigms intended to encourage parallel processing, researchers have relied on intuitions and indirect evidence suggesting that parallel processing is more likely under some conditions than others—for example, with extensive practice (e.g., Hazeltine, Teague, & Ivry, 2002; Hirst, Spelke, Reaves, Caharack, & Neisser, 1980; Schumacher, Seymour, Glass, Fencsik, Lauber, Kieras, & Meyer, 2001; but for bottleneck-based accounts of practice effects, see Ruthruff, Johnston, & Van Selst, 2001, and Ruthruff, Johnston, Van Selst, Whitsell, & Remington, 2003). Others, especially Meyer and Kieras (1999), have determined the conditions under which parallel processing would occur from specific models of processing (see also Logan & Gordon, 2001; Tombu & Jolicœur, 2002).

This article is based on a theoretical analysis of the conditions that determine whether parallel or serial processing is more efficient. In the first section, we present a metatheoretical model of dual-task performance that allows us to assess formally the optimality of serial and parallel processing modes under various circumstances. One surprising implication of this model is that serial processing is almost always more efficient than parallel processing. In light of this implication, repeated demonstrations of seriality do not seem theoretically decisive, because they could result from performance optimization rather than from a structural limitation.

Using our metatheoretical model, we develop an experimental manipulation that can be used to increase the benefit of parallel processing relative to serial processing. In the second section, we present a series of experiments examining the effects of this experimental manipulation on dual-task performance. In general, performance is sensitive to this manipulation

in ways inconsistent with the idea of strict serial processing. Instead, the results demonstrate effects predicted from the idea that participants shift to a more parallel mode of processing when such a mode is more likely to be optimal. In short, the results weaken the claim of an immutable structural bottleneck, as do previous findings that at least some participants tend to shift processing modes in response to instructions emphasizing the use of parallel versus serial strategies (e.g., Schumacher, Seymour, Glass, Fencsik, Lauber, Kieras, & Meyer, 2001).

## A Metatheoretical Framework for Optimization of Dual-Task Performance

We first develop a theoretical framework for analyzing performance in PRP tasks and determining whether a serial or parallel processing mode would be optimal under a given set of conditions. To determine optimal processing, it is necessary to choose a criterion measure to be optimized. One rather compelling performance measure is the total time needed for the performance of both tasks, TRT, measured as the sum of the RTs for the two tasks (i.e., $\mathrm{TRT} = \mathrm{RT1} + \mathrm{RT2}$). It seems appropriate for participants in standard PRP tasks to try to minimize this sum, because they are usually told to make both responses as quickly as possible. We also assume, for simplicity, that task order is fixed, with S1 always presented before S2, as is the case in most PRP studies (for exceptions see, e.g., De Jong, 1995; Pashler, 1990, 1994b; Tombu & Jolicœur, 2002).

To illustrate this theoretical framework, it is helpful to begin with an easily-understood physical analog of the dual-task situation. Consider a car wash staffed by six workers. Suppose that the six workers can either (a) all work together on one car and wash it in two minutes, or (b) split up into two teams of three working in parallel and wash two cars in four minutes. Now suppose that two cars arrive at the car wash almost simultaneously. The driver of each car would like to get his or her car washed as quickly as possible, minimizing the time spent in the car wash. If this is the goal, is it better for the drivers if the six workers (a) work together and wash one car at a time (serial processing mode), or (b) split into two teams and wash both cars at the same time (parallel processing mode)? It is easy to see that the total time spent by the drivers in the car wash—which corresponds to the total RT, $\mathrm{TRT} = \mathrm{RT1} + \mathrm{RT2}$, in the PRP situation—is less

with the serial mode. With this mode the first car is finished after RT1 = two minutes, and the second car is finished after RT2 = four minutes. The two drivers thus spend a total of TRT = six minutes at the car wash. With the parallel mode, however, both cars are finished after RT1 = RT2 = four minutes, so the two drivers spend a total of TRT = eight minutes at the car wash. Clearly, TRT is smaller with the serial mode than with the parallel one.

As this analogy illustrates, serial processing may be more efficient than parallel processing in terms of the overall time needed for the completion of two tasks, even when there is no structural limitation preventing the two tasks from being processed simultaneously. Although this conclusion appears to be counterintuitive, it has also previously been established within the area of scheduling theory (e.g., Conway, Maxwell, & Miller, 1967; Schweickert & Boggs, 1984). Note further that the advantage for serial processing might be even larger if the two cars arrive sequentially rather than simultaneously. If the workers at the car wash divide into two teams and only one car arrives, then half of the workers might stand around rather unproductively until the second car arrives.

Figure 1 illustrates the basic concepts of a more general and formal model subsuming the car-wash analogy as a special case. For simplicity we assume that the processing of each task takes a certain amount of time, $X_s$ or $X_p$, depending on whether the tasks are processed in a serial or parallel mode. In the car-wash analogy we assumed that the two tasks have equal priority and that $X_p = 2 \cdot X_s$, but in the more general case these restrictions need not apply. In keeping with virtually all models that allow limited-capacity parallel processing, however, we do assume a longer processing times for both tasks when they are processed in parallel than when they are processed in serial, so $X_p > X_s$. For simplicity, in the introduction we will develop the metatheoretical model under the additional assumptions that processing time is determined by a single processing stage, is constant across trials, and is equal for the two tasks. Appendices A, B, and C show that these additional assumptions, however, are not essential for the conclusions that we reach. First, Appendix A shows that the same basic conclusions can be reached from a more elaborate model in which tasks are carried out by a sequence of three stages with randomly varying durations. Second, Appendix B addresses the more complex situation in which two tasks are processed in parallel but with unequal emphasis. The hypothesis of parallel processing

includes a whole range of possibilities varying in the relative emphasis on the two tasks (e.g., relative capacity allocations; Navon & Gopher, 1979), and this appendix shows how capacity should be optimally allocated in order to minimize total processing time, TRT. Third, Appendix C extends the conclusions to a more flexible parallel model in which central capacity is reallocated instantaneously according to task demand so that it is always fully used (e.g., Ruthruff, Pashler, & Hazeltine, 2003; Tombu & Jolicœur, 2003).

For the simpler case considered in the introduction, the upper two panels of Figure 1 depict processing with the serial mode, and the lower two panels depict processing with the parallel mode. To relate the framework directly to the PRP paradigm, each mode is illustrated both for a short SOA, on the left side of the figure, and for a long SOA, on the right side. The figure illustrates the limiting cases in which the short SOA is zero and the long SOA is longer than the time needed for Task 1 processing.

---

Insert Figure 1 about here

---

First, consider the serial mode at the short SOA. In this case Task 1 is processed first, and Task 2 must wait until processing of Task 1 has finished.[3] If each task consumes $X_s$ time units, the overall task time (i.e., TRT = RT1 + RT2) is equal to TRT $= 3 \cdot X_s$. Now consider the parallel mode at the short SOA. In this situation, both tasks finish after $X_p$ time units. Hence, the overall task time under this mode is TRT $= 2 \cdot X_p$. Based on this very simple conception, we can already consider whether the serial or parallel mode is more efficient for this case. Clearly, the overall task time is less for the parallel mode than for the serial mode if $2 \cdot X_p < 3 \cdot X_s$; that is, when $X_p < 1.5 \cdot X_s$. In contrast, the overall task time is less for the serial mode than for the parallel mode if $X_p > 1.5 \cdot X_s$, and the two modes yield identical overall task times if $X_p = 1.5 \cdot X_s$.

As an example, consider a typical capacity model in which processing rate increases with the proportion of resources allocated (e.g., Tombu & Jolicœur, 2003). If capacity is divided equally between Tasks 1 and 2, then $X_p = 2 \cdot X_s$, because it takes twice as long to do a given amount of work with only half of the available resources. This was also the situation in the car wash example discussed earlier, because it took twice as long to wash a car when the team had

half as many workers. For such situations, the model illustrated in Figure 1 indicates that an equal division of resources would definitely be suboptimal. Assuming that participants try to optimize performance, then, we would not expect to see parallel processing in such situations even if there were no structural limitation preventing it (i.e., no indivisible bottleneck stage).

Consider next the two panels on the right of Figure 1, which illustrate the consequences of serial and parallel modes at a long SOA. With serial processing, as shown in the upper panel, Task 2 need not be postponed, and the overall task time is $\text{TRT} = 2 \cdot X_s$. With parallel processing, as shown in the lower panel, the participant is prepared to process both tasks simultaneously, so the processing times correspond to the slower parallel mode. The overall task time is thus $\text{TRT} = 2 \cdot X_p$, just as it was with the short SOA. Given our assumption that $X_s < X_p$, the serial mode must always be better than the parallel one at the long SOA.[4] The same conclusion was reached in the car wash example, because of the fact that one team of workers would be idle until the second car arrived.

Two main points emerge from the model illustrated in Figure 1. First, the relative efficiency of the serial and parallel modes depends a great deal on the exact relation between the processing times needed under the two modes, $X_s$ and $X_p$. The parallel mode can only be more efficient than the serial one if $X_p$ is not too much larger than $X_s$. In the car wash example, parallel processing is only optimal if three workers can wash a car in less than 1.5 times as long as six workers (i.e., $X_p < 1.5 \cdot X_s$). Second, the relative efficiency of the two modes also depends on SOA. The parallel mode may be more efficient than the serial one at a short SOA, but it can never be more efficient at a long one. At the long SOA, each task is processed more slowly in the parallel mode than it would be with the serial mode, yet there is no opportunity for gains from parallel processing given that the long SOA prevents task overlap.

Note also that the metatheoretical framework illustrated in Figure 1 is general enough to apply to parallel models other than the capacity models that we commonly use as an illustration of this class. That is, the framework applies to any model in which processing time is increased for parallel processing relative to serial processing, whether the increase is due to capacity limitations or some other type of interference (e.g., outcome conflict; Navon, 1984). Thus, although we will illustrate the discussion of parallel models mainly using capacity models, the

conclusions apply to a broad class of parallel models.

Returning to the question of how researchers might encourage parallel processing, this analysis suggests that one possibility is to manipulate the relative frequencies of different SOAs. Suppose that an experimenter uses just two SOAs and arranges the trials so that the short and long SOAs occur with probabilities of $g$ and $1-g$ respectively. Furthermore, suppose that the participant adopts the parallel mode with probability $c$ and the serial mode with probability $1-c$. Given that the SOAs vary randomly from trial to trial and that the participant must choose a processing mode before the trial starts, the SOA and processing mode must be independent [e.g., $\Pr(\text{short SOA} \cap \text{parallel}) = \Pr(\text{short SOA}) \cdot \Pr(\text{parallel})$]. It is not difficult to compute the average overall task time under these assumptions:

$$\begin{aligned}
\mathrm{E}[\mathrm{TRT}] &= \mathrm{E}[\mathrm{TRT}|\text{short SOA} \cap \text{parallel}] \cdot \Pr(\text{short SOA} \cap \text{parallel}) + \\
&\quad \mathrm{E}[\mathrm{TRT}|\text{long SOA} \cap \text{parallel}] \cdot \Pr(\text{long SOA} \cap \text{parallel}) + \\
&\quad \mathrm{E}[\mathrm{TRT}|\text{short SOA} \cap \text{serial}] \cdot \Pr(\text{short SOA} \cap \text{serial}) + \\
&\quad \mathrm{E}[\mathrm{TRT}|\text{long SOA} \cap \text{serial}] \cdot \Pr(\text{long SOA} \cap \text{serial}) \\
&= c \cdot g \cdot 2 \cdot X_p + c \cdot (1-g) \cdot 2 \cdot X_p + (1-c) \cdot g \cdot 3 \cdot X_s + (1-c) \cdot (1-g) \cdot 2 \cdot X_s \\
&= [2 \cdot X_p - X_s \cdot (g+2)] \cdot c + (2+g) \cdot X_s \qquad (1)
\end{aligned}$$

Thus, $\mathrm{E}[\mathrm{TRT}]$ is a linear function of $c$ with slope $[2 \cdot X_p - X_s\,(g+2)]$. Note that the participant should try to adjust $c$, which varies between 0 and 1, to minimize $\mathrm{E}[\mathrm{TRT}]$.

Because this function is predicted to be linear, it follows that the optimal setting of $c$ must be either 0 or 1, but cannot be in between. The optimal setting is $c = 1$ (i.e., always parallel processing) when the slope is negative, because in this case TRT decreases as $c$ increases. Thus, the parallel mode is best when

$$\begin{aligned}
0 &> [2 \cdot X_p - X_s\,(g+2)]\,, \\
\frac{X_p}{X_s} &< \frac{g+2}{2}. \qquad (2)
\end{aligned}$$

Thus, parallel processing is optimal when the ratio of parallel to serial processing times is relatively small, and it needs to be especially small when the proportion of trials with short SOAs is small. Conversely, the optimal setting is $c = 0$ (i.e., always serial processing) when the slope is

positive, which occurs when

$$\frac{X_p}{X_s} > \frac{g+2}{2}. \tag{3}$$

Now $g$ is the only parameter over which the experimenter has direct experimental control. The question, then, is how large $g$ has to be for the parallel mode to be more efficient than the serial one. The answer can be obtained by rearranging inequality 2, which yields

$$g > 2 \cdot \left( \frac{X_p}{X_s} - 1 \right). \tag{4}$$

Thus, an experimenter wishing to encourage parallel processing should attempt to use a value of $g$ that is at least as large as the right side of Inequality 4. If $X_p/X_s = 1.33$, for example, $g$ should be at least .66 for the parallel mode to be optimal. It is of course difficult to determine the exact value needed for $g$ in practice, because the values of $X_p$ and $X_s$ are unknown. In fact, there is no guarantee of eliciting parallel processing even with $g = 1$, because serial processing will always be better than parallel processing even in that case if $X_p > 1.5 \cdot X_s$. In the absence of independent information about the relative processing times $X_p$ and $X_s$, however, experimenters wishing to elicit parallel processing can only increase $g$ and hope that the processing times are such that parallel processing is optimal for at least some conditions or participants.

In summary, two main conclusions emerge from this analysis (cf. Table 1). First, under many conditions (e.g., $X_p > 1.5 \cdot X_s$) the serial mode is simply more efficient than the parallel one, contrary to the common intuition that tasks get done more rapidly when they are done simultaneously. This conclusion is perhaps surprising given the massively parallel anatomical structure of the brain (e.g., Ghez, 1991; Hubel, 1979; Kandel, 1991; Martin, 1991; Rauschecker, 1998; Rumelhart & McClelland, 1986; Wässle, Grunert, Rohrenbeck, & Boygott, 1990), although this massively parallel structure has been convincingly demonstrated only for sensory input processes (cf. Miller & Ulrich, 2003), not for central response selection processes involved in performing arbitrary choice RT tasks. It follows from this conclusion that participants attempting to optimize their behavior—in accordance with the instructions—would often adopt the serial processing mode. In particular, the present analysis shows that serial processing should be preferred when parallel processing is not particularly fast (i.e., $X_p/X_s > 1.5$), as shown in the upper half of the table. Second, when parallel processing is fast enough that it may be more

efficient than serial processing (i.e., when $1 < X_p/X_s < 1.5$), the parallel mode still only provides gains when short SOAs are relatively frequent (lower half of the table). This clearly suggests that experimenters attempting to document parallel processing should use a preponderance of short SOAs (e.g., Schumacher et al., 2001).

---

Insert Table 1 about here

---

**Present Experiments**

In the present experiments we manipulated the probabilities of different SOAs within different blocks of trials. As shown in Table 2, for example, in some blocks short SOAs were relatively frequent and long SOAs were relatively infrequent, whereas in other blocks the reverse was true. According to the optimization framework given in the introduction, this manipulation could influence the relative efficiencies of the serial and parallel processing modes.

---

Insert Table 2 about here

---

If participants can reduce their total RT by processing in the parallel mode, then they should tend to do that more when the short SOA is likely (e.g., condition SF of Table 2) than when the long SOA is likely (e.g., condition LF). Figure 2 illustrates the changes in RT1 and RT2 predicted by this shift in processing mode. In general, RT1 should be longer when the short SOA is likely, because Task 1 processing is slower in the parallel mode than in the serial mode. For RT2, the model predicts an interaction between the actual SOA and the likely SOA. At short SOAs, RT2 should be smaller when the short SOA is likely than when the long SOA is likely, partly because the waiting period is avoided by parallel processing. At long SOAs, however, RT2 should be smaller when the long SOA is likely than when the short one is, because the serial mode is inherently more efficient when there is no task overlap.

Insert Figure 2 about here

In contrast, what would the bottleneck model predict for this experiment? In its modal form, the model predicts that neither RT1 nor RT2 would depend on the relative likelihoods of the various SOAs. After all, in this model processing is always serial. The time needed for Task 1 should not depend on either the actual SOA or the likely SOA, because this task is always processed first. The time needed for Task 2 should depend on the actual SOA, of course, because this SOA influences the time spent waiting for the bottleneck process, which inflates RT2. Assuming that nothing happens during the waiting period, however, RT2 should not depend on whether the waiting period is usually long or usually short, because there is simply no processing of Task 2 during that period.

## Experiment 1

Experiment 1 employed the same PRP tasks used by Pashler (1994b). Participants performed two choice-RT tasks on each trial. The stimulus for Task 1 was a tone of low or high frequency, and the correct response was to press a button with the middle or index finger of the left hand. The stimulus for Task 2 was a letter H or O, to which the correct response was to press a button with the middle or index finger of the right hand. S1 was always presented first, and the SOA from the tone to the letter ranged from 16 to 1,000 ms.

There were two different types of trial blocks (Table 2). In one block, trials with short SOAs were more frequent than trials with long SOAs (Condition SF). In contrast, the other block was comprised of more trials with long SOAs than with short ones (Condition LF). The main purpose of the experiment was to test whether this manipulation of the SOA distribution would influence the processing mode. Specifically, given the theoretical framework outlined in the introduction, the condition with frequent short SOAs should provide a greater opportunity for parallel processing if $X_p < 1.5 \cdot X_s$ for at least some participants. In contrast, if long SOAs occur more frequently than short ones, participants should tend to process the two tasks serially. If participants do tend to shift processing modes in accordance with this model, the observed RTs

should reveal a pattern analogous to that depicted in Figure 2. Specifically, mean RT1 should be larger in condition SF than in condition LF, and the function relating RT2 to SOA should have a shallower slope in condition SF than in condition LF.

*Method*

*Participants.* Twenty-five students at the University of Tübingen participated in a single session in return for partial fulfillment of a curriculum requirement or EURO 14. The data of five participants had to be discarded because they tended to group their responses on most trials. Results are thus reported for 20 participants (15 female) with an average age of 23.3 years.

*Apparatus and stimuli.* The visual stimulus was a white letter (H or O) presented centrally on a computer monitor with a resolution of 640 by 480 pixels. Each letter was 1.6 cm wide and 2 cm high and subtended visual angles of $1.83^o$ and $2.29^o$ from a typical viewing distance of 50 cm. The background color of the monitor was dark blue. The auditory stimulus was a tone of either 300 or 900 Hz presented through the speakers of the PC at approximately 70 dB. Each stimulus lasted 200 ms. The stimulus onsets were synchronized with the refresh rate of the monitor.

Participants responded on external response buttons with a high temporal registration accuracy. There was a separate panel with two response buttons for each hand. The two buttons on each panel were separated by 2.5 cm. A force of approximately 150 cN was required to register a response. Both forearms of the participant rested comfortably on a table, and the response fingers rested on the respective response buttons.

*Design.* The SF and LF conditions were administered in separate blocks of 480 trials each. The numbers of trials per SOA in each condition are shown in Table 2. Each experimental condition was preceded by a practice block of 80 trials with the same proportion of trials at each SOA value as in the subsequent experimental block. Half of the participants performed condition SF before condition LF; the other half performed LF before SF. At each SOA all possible factorial combinations of S1 and S2 were tested equally often.

*Procedure.* Participants were given written instructions describing the tasks and instructing participants to respond as quickly and accurately as possible to each stimulus. Participants were

also instructed to place equal emphasis on both tasks.[5]

Each trial began with the presentation of a plus sign as a fixation point at the center of the display for 1 s. When it disappeared there was a fixed foreperiod duration of 0.5 s. After the foreperiod had elapsed, the tone stimulus S1 was presented, and after the corresponding SOA had elapsed, the visual stimulus S2 was presented. Participants responded to the tone stimulus with their left hands, pressing a button with the middle finger for low tones and with the index finger for high tones. They responded to S2 with their right hands, pressing a button with the index finger for the letter H and with the middle finger for the letter O. The message "Fehler!" (Error!) was presented for 1 s at the end of a trial if an error was made in either task. The intertrial interval between the offset of the error message and the onset of the fixation point of the next trial was 2.5 s. However, when no error was made the intertrial interval between the second response and the onset of the fixation point was 1.5 s.

After each set of 40 trials, there was a pause of 20 s. During this pause the participant received feedback about the overall RT and the percentage of errors for the preceding 40 trials. After the 20 s had elapsed, a message on the screen asked the participant to initiate the next set of 40 trials by pressing one of the response buttons. When the first condition had been completed (i.e., following 560 trials = 80 practice trials + 480 experimental trials), participants were asked to pause for 5 to 10 min. The complete session lasted approximately 1 h 50 min.

*Results*

The first 80 trials with each SOA distribution (i.e., SF or LF) were considered practice and thus excluded from the analysis. Before analyzing the RT results of the experimental blocks, we screened all trials in these blocks for response errors. The overall percentage of response errors was 4.6%, and such trials were excluded from RT analysis. In a next step, all trials with correct responses were screened for RTs less than 150 ms or greater than 3,000 ms. The overall percentages of such trials were 0.13% and 0.0%, respectively, and they were also excluded from the computations involving RTs. Finally, in the main analyses we screened the data for trials in which the interresponse interval, IRI, was less than or equal to 100 ms. These trials are of special interest because they could result from a strategy of response grouping (e.g., Borger, 1963) that

could introduce unwanted effects into the RT data. The percentages of trials with such small IRIs are reported separately for each condition below. Only the remaining trials (i.e., approximately 89% of all trials in the experimental blocks) were included in the analyses of RT1 and RT2. Other analyses were also conducted with larger and smaller IRI cutoffs (i.e., IRI=50 ms and IRI=140 ms), but the results of theses additional analyses will not be reported because they were virtually identical to those with the 100 ms cutoff.

Figure 3 depicts the average values of the dependent variables as a function of SOA distribution and SOA. Analyses of variance (ANOVAs) were computed for each dependent variable separately, using within-subjects factors of SOA (16, 133, 500, vs. 1,000 ms) and SOA distribution (SF vs. LF).

---

Insert Figure 3 about here

---

*Reaction times.* In the ANOVA on RT1, the main effect of SOA was not significant, $F(3, 57) = 1.92$, $MSE = 10,327$, $p > .1$. As predicted by the optimization framework, mean RT1 was numerically larger in condition SF than in LF, although this effect only approached statistical significance, $F(1, 19) = 3.69$, $MSE = 11,222$, $p < .1$. The interaction of SOA and SOA distribution was highly significant, however, $F(3, 57) = 7.66$, $MSE = 3,007$, $p < .001$. Inspection of the means in Figure 3 suggests that this interaction arose because of an overall tendency for RT1 to increase at the least frequent SOAs. Consistent with this impression, a linear trend analysis indicated that RT1 increased with SOA in the SF condition, $p < .01$, whereas it tended to decrease with SOA in the LF condition, although the latter tendency was not reliable, $p > .2$. Thus, it appears that performance of Task 1 may be disrupted somewhat when an unlikely SOA is used. This interaction is not predicted either by bottleneck models or by the optimization framework, but it might be explained in terms of perceptual interactions within either context. Specifically, suppose that the arrival of S2 tends to interfere with Task 1 processing (e.g., Jolicœur & Dell'Acqua, 1999; Wühr & Müsseler, 2002), and that this interference is especially large when S2 is unexpected. In that case there would be especially large interference, and thus especially large RT1 values, at short SOAs in the LF condition.

In the parallel ANOVA on RT2, this measure increased substantially as SOA decreased, $F(3, 57) = 291.12$, $MSE = 6,073$, $p < .001$, reflecting the usual PRP effect. Although the main effect of SOA distribution on RT2 was not significant, $F(1, 19) = 0.17$, $MSE = 11,990$, $p > .6$, there was a highly significant interaction of SOA distribution and SOA, $F(3, 57) = 6.04$, $MSE = 2,806$, $p < .002$. In qualitative agreement with the predictions of the optimization framework (cf. Figure 2), mean RT2 at short SOAs tended to be less in condition SF than in conditions LF, whereas the reverse was true at long SOA values.

A further analysis was conducted to examine in detail whether the function relating RT2 to SOA was steeper in condition LF than in condition SF, as predicted by the optimization framework. In a first step, the slope $m = \Delta RT/\Delta SOA$ was determined for each of the three segments of the SOA-RT2 function [i.e., $m_1 = (\mathrm{RT2}_{133} - \mathrm{RT2}_{16})/(133 - 16)$, $m_2 = (\mathrm{RT2}_{500} - \mathrm{RT2}_{133})/(500 - 133)$, and $m_3 = (\mathrm{RT2}_{1,000} - \mathrm{RT2}_{500})/(1,000 - 500)$]. In a second step, we averaged these three slopes to obtain a measure of the overall steepness of the SOA function. This computation was performed for each participant, and the resulting values were averaged across participants. The overall average slopes were -0.74 and -0.58 for conditions LF and SF, respectively, and these differed significantly according to a one-sided $t$-test for matched pairs, $t(19) = 2.34$, $p < .02$. Thus, this additional analysis strengthens the claim that the distribution of SOA values modulates the steepness of the function relating RT2 to SOA. In particular, the function was steeper when long SOAs were frequent than when short SOAs were frequent. A similar analysis on the average of the first two segments also revealed a significant difference in slopes (SF= $-0.81$, LF= $-1.02$), $t(19) = 2.07$, $p < .05$. For the first segment considered in isolation, the test just reached statistical significance (SF= $-1.01$, LF= $-1.35$), $t(19) = 1.69$, $p < .05$.

*Percentages of correct responses and of responses with short IRIs.* The percentage of correct responses varied slightly yet almost significantly with SOA, $F(3, 57) = 2.43$, $MSE = 6.8$, $p < .1$, with the lowest accuracy at the shortest SOA and the highest accuracy at the longest SOA. The main effect of SOA distribution was not significant, $F(1, 19) = 2.50$, $MSE = 9.2$, $p > .1$, and the interaction of the two factors was marginally significant, $F(3, 57) = 2.53$, $MSE = 4.6$,

$.05 < p < .1$.

Overall, 6.5% of trials with correct responses had IRIs less than 100 ms. Small IRIs were more frequent in condition SF than in condition LF, $F(1, 19) = 7.41$, $MSE = 186.0$, $p < .02$. Consistent with previous reports of especially prevalent response grouping at short SOAs (e.g., Ivry, Franz, Kingstone, & Johnston, 1998; Lien, Schweickert, & Proctor, 2003; Pashler & Johnston, 1989; Ruthruff, Pashler, & Hazeltine, 2003), the percentage of small IRIs decreased from 13.2% to 0.4% as SOA increased from 16 to 1,000 ms, $F(3, 57) = 13.41$, $MSE = 113.9$, $p < .001$, and this effect was especially large in condition SF, where small IRIs were more prevalent anyway, $F(3, 57) = 5.71$, $MSE = 40.7$, $p < .01$.

*Correlation of RT1 and RT2.* The bottleneck model predicts strong positive correlations between RT1 and RT2, especially at short SOAs, and this prediction has been confirmed repeatedly (e.g., Pashler & Johnston, 1989; Welford, 1967). Unfortunately, the presence of such strong positive correlations does not selectively support bottleneck models, because alternative models can also accommodate them (e.g., Navon & Miller, 2002; Tombu & Jolicœur, 2003). Moreover, we were unable to derive any clear predictions about the effects of SOA distribution on correlations from either bottleneck models or from the optimization framework. Nonetheless, we report these correlations for completeness and for their possible relevance to future theoretical efforts.

The bottom panel in Figure 3 shows the correlation between RT1 and RT2 as a function of SOA for each SOA distribution. The correlation coefficient between RT1 and RT2 was computed across trials for each experimental condition and for each participant, and the figure shows the average of these values across participants. Consistent with previous research, RT1 and RT2 were correlated positively, and the correlation decreased as SOA increased, $F(3, 57) = 68.90$, $MSE = 0.024$, $p < .001$. Interestingly, the correlation coefficient was larger in SF than in LF. Although this effect was only marginally significant, $F(1, 19) = 3.06$, $MSE = 0.007$, $.05 < p < .1$, the difference became significant at the longest SOA as indicated by the significant interaction of the two factors, $F(3, 57) = 5.21$, $MSE = 0.0123$, $p < .001$.

*Discussion*

Two main results shown in the top panel of Figure 3 provide support for the thesis that the distribution of SOAs affects performance in PRP designs, as predicted by our optimization framework. First, the effect of SOA on RT2 was smaller when short SOAs were frequent than when long SOAs were frequent. This is in accordance with the facts that (a) the parallel processing mode tends to be relatively efficient when short SOAs are common, and (b) RT2 tends to be affected less by SOA with the parallel mode than with the serial mode. Second, RT1 tended to be longer when short SOAs were frequent than when long SOAs were frequent, although this effect only approached significance.

Although the range and distribution of SOAs would be expected to influence the use of a response grouping strategy (e.g., Pashler, 1994b), further analyses suggested that response grouping did not contribute in any important way to the critical results observed here. For one thing, the patterns of RT1 and RT2 means were hardly affected by exclusion of trials with IRIs less than 50 versus 100 ms. For another, the strongest result of this experiment involved RT2 (i.e., the interaction of SOA and SOA distribution). Response grouping influences mainly RT1 rather than RT2 (e.g., Pashler & Johnston, 1989), so it is unlikely that grouping would be responsible for this result. Perhaps most importantly, additional median-split analyses were carried out to examine the results separately for trials with IRIs shorter versus longer than the median IRI (cf. Hommel, 1998).[6] In the analysis of RT1, the key effect of SOA distribution was not significantly different for trials with short versus long IRIs ($p > .5$). In the analysis of RT2, the key interaction of SOA and SOA distribution was also statistically independent of IRI ($p > .7$). Under the standard assumption that trials with grouped responses should have relatively short IRIs, the fact that these key effects are not modulated by IRI provides strong evidence that they are not an artifact of response grouping.

Figure 3 illustrates one other interesting difference between the conditions with short and long SOAs frequent: There were more trials with short IRIs when short SOAs were frequent. This finding is also consistent with the idea that participants adopt a more parallel processing mode when a short SOA is likely, because parallel processing tends to produce shorter IRIs than does

serial processing. In the serial mode, queueing makes it virtually impossible for processing of Task 2 to catch up with processing of Task 1, so small IRIs are unlikely unless an explicit strategy of response grouping is used. For example, within the standard three-stage serial model considered in Appendix A, the IRI must be at least as great as the time needed for processing of Task 2 within the bottleneck stage (i.e., $B2_s$), as long as the motor times for the two tasks are equal (cf. Figure A1). In contrast, IRIs can be much smaller in the parallel processing mode, because there is no enforced queueing and the two tasks can in principle finish at the same time.

The increased number of small IRIs in the SF condition is consistent with the hypothesis that processing is more often parallel in that condition. An alternative explanation of this increase, however, is that it simply reflects a stronger tendency to group responses when the stimuli tend to occur near-simultaneously than when they tend to occur with a large temporal separation. To evaluate this account of the increase, we compared the full frequency distributions of IRIs for conditions SF versus LF, pooling across participants, as shown in Figure 4. It is evident that the effect of SOA distribution was not simply to increase the number of very small IRIs, but that instead this effect was present throughout most of the IRI range. Thus, we conclude that IRIs do tend to be smaller in the SF condition than in the LF condition, in a manner at least qualitatively consistent with what would be expected from a greater tendency toward parallel processing in the former condition.[7]

---

Insert Figure 4 about here

---

In summary, the results of Experiment 1 support the theoretical ideas that people can make some adjustments in the extent to which they use serial versus parallel processing modes, and that they do so partly in response to the distribution of SOAs within a block. These ideas are of course more consistent with the predictions of the optimization framework than with those of the bottleneck model, and they thereby tend to support the position that the serial mode of processing may be merely an effective strategy—not a biological constraint—in PRP tasks. Before drawing firm theoretical conclusions from the results, however, it seems appropriate to replicate them and investigate their generality.

## Experiment 2

Experiment 2 was identical to Experiment 1 except that it employed more skewed distributions of SOAs, in an attempt to magnify the effect of SOA distribution and thereby increase the effects observed in Experiment 1. Specifically, Experiment 2 had even higher percentages of very short or very long SOAs.

### *Method*

*Participants.* A fresh sample of 24 students (17 female) participated in a single session. Their average age was 28.3 years. The data of four participants were excluded from the data analyses because these participants grouped their responses excessively.

*Apparatus and stimuli.* The apparatus and stimuli were identical to those used in Experiment 1.

*Procedure and design.* The design and procedure were identical to those in Experiment 1 except that the SOA distributions in conditions SF and LF were more extremely skewed. Specifically, condition SF included 336, 48, 48, and 48 trials with SOAs of 16, 133, 500, and 1,000 ms, respectively. In contrast, condition LF included 48, 48, 48, and 336 trials, respectively, for these SOA values.

### *Results*

The data were subjected to the same analyses as in Experiment 1. Response errors occurred in 5.8% of all trials, and these trials were excluded from the analyses of RTs. The percentages of RT outliers were again small and virtually identical to those in Experiment 1 (i.e., 0.1% of trials with RTs less than 150 ms, 0.0% of trials with RTs larger than 3,000 ms). As in Experiment 1, trials with such outliers and with IRIs less than 100 ms were also excluded from the analyses of RT1 and RT2. In total, then, approximately 11% of all trials were excluded from these analyses. The main average results are presented in Figure 5, and each dependent variable was again analyzed with an ANOVA including factors of SOA distribution and SOA.

Insert Figure 5 about here

*Reaction times.* Once again, RT1 was shorter in condition LF than SF. In contrast to Experiment 1, however, this effect was highly reliable, $F(1, 19) = 12.59$, $MSE = 27,379$, $p < .01$, as was the main effect of SOA, $F(3, 57) = 8.15$, $MSE = 4,647$, $p < .001$. Moreover, the two factors again produced a significant interaction, $F(3, 57) = 4.15$, $MSE = 4,700$, $p < .01$. As is obvious from Figure 5, the RT1 difference between LF and SF increased as SOA was lengthened.

In the parallel analysis of RT2, the SOA factor again produced a strong main effect, $F(3, 57) = 217.29$, $MSE = 8,027$, $p < .001$. As in Experiment 1, the main effect of SOA distribution was not significant, $F(1, 19) = 0.52$, $MSE = 23,136.7$, $p > .4$, but there was again a strong interaction between this distribution and SOA, $F(3, 57) = 5.31$, $MSE = 2,806$, $p < .01$. Once again, mean RT2 decreased more rapidly with SOA in the LF condition than in the SF condition, as indicated by slope analyses identical to those used in Experiment 1. The average slopes over all three SOA segments were -0.74 for condition LF and -0.63 for condition SF, and these values were reliably different, $t = 2.40$, $df = 19$, $p < .02$. The same effect was significant in a comparison of the average slopes of the first two segments (i.e., SOAs from 16 ms to 500 ms: SF= $-0.86$, LF= $-1.01$), $t(19) = 2.18$, $p < .025$, and marginally significant in a comparison of slopes involving only the first segment (i.e., SOAs of 16 ms vs. 133 ms: SF= $-1.17$, LF= $-1.34$), $t(19) = 1.69$, $p < .1$.

*Percentages of correct responses and of responses with short IRIs.* As in Experiment 1, participants produced somewhat fewer correct responses at short SOAs than at long ones, $F(3, 57) = 3.30$, $MSE = 13.4$, $p < .05$. In addition, and in contrast to Experiment 1, fewer correct responses occurred in condition LF than in SF on average across SOAs, $F(1, 19) = 6.06$, $MSE = 9.8$, $p < .025$. The interaction of these two factors was not significant.

The overall percentage of responses with IRI $< 100$ ms was 5.6%, and thus was virtually identical to the overall figure obtained in Experiment 1. Also consistent with the results of Experiment 1, this percentage decreased from 10.2% to 0.8% as SOA increased, $F(3, 57) = 3.32$,

$MSE = 298.8$, $p < .05$. Interestingly and in marked contrast to the previous experiment, neither the effect of SOA distribution nor its interaction with SOA was significant, $F$s$< 1$. As is shown in Figure 6, IRIs again tended to be smaller in condition SF than in condition LF at all SOAs, but the size of this effect was smaller than was obtained in Experiment 1, especially at the short SOAs. This change may have resulted from a reduced tendency to group responses in this experiment.

---

Insert Figure 6 about here

---

*Trial-to-trial correlation between RT1 and RT2.* As is evident in the bottom panel of Figure 5, the pattern of correlations between RT1 and RT2 basically replicated that found in Experiment 1. Once again, the correlation decreased as SOA increased, $F(3, 57) = 100.98$, $MSE = 0.024$, $p < .001$. Theoretically more interesting and in agreement with Experiment 1, the correlation was larger in condition SF than in LF. This time, however, this main effect was highly significant, $F(1, 19) = 9.47$, $MSE = 0.036$, $p < .01$. Although the effect of SOA condition increased numerically as before with SOA, the interaction of these two factors did not reach statistical significance this time, $F(3, 57) = 1.59$, $MSE = 0.024$, $p > .2$.

*Discussion*

The results of this experiment demonstrate even stronger effects of the distribution of SOAs than did Experiment 1, at least for RT1, presumably because of the stronger manipulation of this distribution. As in Experiment 1, one critical finding was the larger effect of SOA on RT2 when long SOAs were frequent than when short SOAs were frequent. Unlike Experiment 1, though, this experiment produced a second highly significant effect of SOA distribution: average values of RT1 were smaller when long SOAs were frequent than when short ones were, consistent with the use of a more serial processing mode in that case. As in Experiment 1, it does not appear that these results are attributable to response grouping, because they are unchanged across various reasonable IRI cutoffs for grouped responses. In addition, median-split analyses again show the same key results for trials with IRIs smaller versus larger than the median. Neither the effect of SOA distribution on RT1 nor the interaction of SOA and SOA distribution on RT2 was

significantly modulated by large versus small IRI status ($p > .5$ and $p > .25$, respectively).

Figure 5 illustrates two other interesting differences between the SF and LF conditions that presumably emerged because of the stronger distribution manipulation (i.e., more skewed SOA distributions). First, as in Experiment 1, IRIs tended to be smaller when short SOAs were frequent than when long ones were, consistent with a shift towards a parallel mode in the former case. Second, RT1 and RT2 were more strongly correlated when short SOAs were frequent. The implications of the latter result are not entirely clear, because RT1/RT2 correlations may be sensitive to a number of factors and can be fairly high within parallel processing models as well as serial ones (Navon & Miller, 2002). One possible interpretation, however, is that correlations tend to be higher with parallel processing because two tasks carried out at the same time tend to be affected in the same way by moment-to-moment fluctuations relevant to task performance (e.g., arousal).

## Experiment 3

In the first two experiments, each task required a manual response. It is possible, however, that structural interference at a motor level might promote serial rather than parallel processing processing when both tasks use the same response modality (e.g., Allport, 1980; De Jong, 1993; Keele, 1973; McLeod, 1977, 1978, 1980; Wickens, 1976; but see Pashler, 1990 for a different view). This interference might be difficult to overcome, and it might cause participants to process in a serial mode even though the manipulation of SOA distribution favors parallel processing at a central level. If so, it might be easier to encourage parallel processing—and to obtain larger effects of the distribution of SOAs—in an experiment without structural interference at the motor level.

In an attempt to reduce the potential motor interference, Experiment 3 employed different response modalities for the two tasks to assess whether stronger signs of parallel processing would be obtained in such a situation. Specifically, Task 1 required a manual response and Task 2 required a vocal one. Thus, Experiment 3 examined whether the previous effects would generalize or even be magnified when the two tasks used different response modalities.

*Method*

*Participants.* Twenty students (13 female) participated in a single session. Their average age was 27.0 years.

*Apparatus, stimuli, and responses.* Based on the results of pilot testing, the sensory modalities of S1 and S2 were reversed relative to the previous experiments in order to maximize mean IRI and minimize response grouping. The Task-1 stimulus was a single letter presented visually. The two letter alternatives were Q and M, presented in the same font used in the previous experiments, and these letters were chosen to minimize interference with the vocal responses required for Task 2. Participants responded with the left index finger to the letter Q and with the right index finger to the letter M.

The stimulus for Task 2 was a tone of either 200 or 1,200 Hz, presented for the same duration and intensity as in the previous experiments. A greater frequency difference was used to facilitate tone discrimination. The two vocal response alternatives for this task were "tief" and "hoch", the German words for low and high, respectively. The participant's speech signal was registered by a microphone, which was amplified by a voice-key (Rieder, Germany). As soon as the amplitude of the speech signal exceeded a fixed threshold, the voice-key produced a digital output signal registered by the PC as the vocal RT. The threshold value was adjusted individually for each participant at the beginning of the session to achieve maximal sensitivity with minimal false alarms. In addition, the speech signal was transmitted via headphones to the experimenter, who sat in another room outside the sound-proofed testing chamber. The experimenter identified incorrect vocal responses so that feedback could be provided at the end of a trial if an error had occurred. In such cases the word "Wortfehler!" (word error) appeared on the screen for 1,000 ms.

*Procedure and design.* The design and procedure were identical to those of the previous experiments except that only three rather than four levels of SOA were employed (i.e., 16, 133, and 1,000 ms). The SF condition included 192, 192, and 96 trials with SOAs of 16, 133, and 1,000 ms, respectively. In contrast, the LF condition included 48, 48, and 384 trials, respectively, at each of these SOA values.

*Results*

The data were subjected to the same analyses as in Experiments 1 and 2. Response errors occurred in 4.4% of all trials, and these trials were excluded from the analyses of RTs. The percentages of RT outliers were again small (i.e., 1.2% of trials with RTs less than 150 ms, 0.0% of trials with RTs larger than 3,000 ms). As in the previous experiments, trials with such outliers and with IRIs less than 100 ms were also excluded from the analyses of RT1 and RT2. In total, then, approximately 7% of all trials were excluded from these analyses. The main average results are presented in Figure 7, and each dependent variable was again analyzed with an ANOVA including factors of SOA distribution and SOA.

---

Insert Figure 7 about here

---

*Reaction times.* The analysis of RT1 indicated that this measure was again significantly shorter in condition LF than SF, $F(1, 19) = 5.86$, $MSE = 1,432$, $p < .05$. It was not significantly affected by SOA ($p > .2$), but the interaction of these two factors was again significant, $F(2, 38) = 3.71$, $MSE = 130$, $p < .05$, despite being rather small numerically. Specifically, the LF condition again had a slightly larger advantage, relative to SF, at the longest SOA.

In the parallel analysis of RT2, SOA had its usual strong main effect, $F(2, 38) = 215.60$, $MSE = 1,707$, $p < .001$. There was again a strong interaction of SOA and SOA distribution, $F(2, 38) = 16.30$, $MSE = 507$, $p < .001$, with a larger decrease in RT2 across SOAs in the LF condition than in the SF condition. Slope analyses yielded values of -0.51 and -0.56 for the SF and LF conditions, respectively, over the SOA range from 16–133 ms, and these values did not differ significantly ($p > .2$). The slopes were significantly different for the SOA range from 133–1000 ms, however, $t(19) = 4.98$, $p < .001$.

*Percentages of correct responses and of responses with short IRIs.* As shown in Figure 7, responses were least accurate at the shortest SOA, $F(2, 38) = 10.20$, $MSE = 17.1$, $p < .01$, but neither the main effect of SOA distribution nor the interaction of this factor with SOA was significant ($p$'s $> .13$).

At 1.6%, the overall percentage of responses with IRI < 100 ms was smaller in this experiment than in the previous two experiments. This measure was affected significantly by SOA, $F(2, 38) = 18.32$, $MSE = 6.84$, $p < .001$, by SOA distribution, $F(1, 19) = 6.27$, $MSE = 11.20$, $p < .025$, and by the interaction of these two factors, $F(2, 38) = 5.13$, $MSE = 4.16$, $p < .02$. In all cases, the directions of these effects were similar to those obtained in the previous experiments. Figure 8 depicts these effects at the distributional level and shows that there are many more IRIs in the 200–300 ms range in condition SF than in condition LF for short SOAs.

---

Insert Figure 8 about here

---

*Trial-to-trial correlation between RT1 and RT2.* The correlations of RT1 and RT2 were distinctly smaller overall than those observed in Experiments 1 and 2, especially at short SOAs, suggesting that the use of different response modalities increased task independence (cf. Rinkenauer, Ulrich, & Wing, 2001). As in the previous experiments, the correlations between RT1 and RT2 decreased as SOA increased, $F(2, 38) = 75.34$, $MSE = 0.021$, $p < .001$. Correlations did not differ across SOA distributions, however, nor was there an interaction of SOA and SOA distribution ($p$'s $> .25$).

*Discussion*

The results of this experiment again show highly reliable effects of the distribution of SOAs. Specifically, compared to blocks in which long SOAs were frequent, blocks in which short SOAs were frequent yielded a larger value of RT1 and a shallower slope relating RT2 to SOA. Thus, this experiment again replicated the two key effects that were predicted directly from the optimization framework and that are difficult to explain in terms of the bottleneck model.

Despite the use of different response modalities for the two tasks, the key signs of parallel processing were not much larger numerically than those obtained in the previous two experiments using manual responses for both tasks. Crudely speaking, then, it appears that the SF condition in this experiment produced approximately the same degree of shift toward more parallel processing as it did in the previous experiments. It should be emphasized, however, that there are

signs that processing was much more parallel in both conditions (SF and LF) in this experiment than in the previous ones. Specifically, in this experiment the slopes of the functions relating RT2 to SOA were much shallower than in the previous experiments, even considering only the two shortest SOAs. Specifically, the slopes here were approximately -0.5, not -1 as predicted by the bottleneck model. Thus, processing may have been much more parallel overall, in condition LF as well as condition SF, due to the use of different response modalities. This pattern is quite consistent with previous suggestions that dual-task interference is greatly reduced when the two tasks use different response modalities (e.g., Allport, 1980 De Jong, 1993; Keele, 1973; McLeod, 1977, 1978, 1980; Wickens, 1976; but see Pashler, 1990).

As in the previous experiments, it does not appear that the evidence of parallel processing was an artifact of response grouping. The results were again unchanged using various reasonable alternative IRI cutoffs for grouped responses, and the increased IRIs relative to the previous experiments meant that there was less grouping in any case. Moreover, median-split analyses again indicated that there was no more evidence of parallel processing in trials with IRIs smaller than the median than in those with larger IRIs. In fact, for this experiment the interaction of SOA and SOA distribution observed with RT2 was significantly larger for trials with long IRIs than for trials with short ones, $F(2, 38) = 3.25$, $MSE = 578$, $p < .05$, which is just the opposite of what would be expected if the interaction were an artifact of grouping.

## General Discussion

We began this article with the assumption that participants in PRP tasks seek to optimize their overall performance by minimizing the total time needed to respond in the two tasks, TRT = RT1 + RT2. The introduction (see also Appendices A, B, and C) explored the consequences of that assumption and established within a fairly general metatheoretical framework that serial rather than parallel processing would be optimal in most cases. Specifically, parallel processing is optimal only when (a) the processing time required under the parallel mode is not much longer than the processing time required under the serial mode (i.e., $X_p < 1.5 \cdot X_s$), and (b) the SOA separating task onsets is usually quite short. Moreover, this is true for a broad class of parallel models, including both limited-capacity models (e.g., Kahneman, 1973) and

outcome conflict models (e.g., Navon, 1984) as special cases.

The conclusion that serial processing is generally optimal in the PRP paradigm led us to question whether evidence of serial processing in this paradigm really implies a structural limitation preventing parallel processing, as entailed by the bottleneck model. Instead, serial processing might emerge mainly from participants' efforts to optimize performance, in keeping with the usual experimental instructions.

Based on this analysis of the optimal processing mode, we conducted three experiments in which we manipulated the frequencies of the different SOAs. Each compared one block in which short SOAs were especially frequent against another block in which long SOAs were especially frequent. As noted above, the optimization framework suggests that participants ought to use a more parallel processing mode when short SOAs are frequent than when long SOAs are frequent. If parallel processing is not possible due to a structural bottleneck, of course, then such optimality considerations could not influence processing mode in any case. Experiments 1 and 2 employed two tasks requiring manual responses, whereas Experiment 3 employed two tasks involving different response modalities.

*Evidence for Adjustments in Processing Mode*

The results indicate that the relative likelihood of short versus long SOAs does affect performance. In particular, there was evidence of more parallel processing in the condition with frequent short SOAs than in the condition with frequent long ones, as was predicted from the optimization framework. One important result is that the function relating RT2 to SOA has a shallower slope when short SOAs are likely than when long SOAs are likely. This decrease in the slope of the function relating RT2 to SOA would be expected if participants tended to prepare for parallel processing when a short SOA was likely but for serial processing when a long one was likely, as they should tend to do in order to minimize TRT. The second important result is that RT1 tended to be longer when short SOAs were likely than when long SOAs were likely. This change in RT1 would also be expected if processing is more parallel when a short SOA is likely, because Task 1 processing should take longer when it is processed in parallel with Task 2 than when it is processed by itself in a serial mode. Both of these results were obtained whether the

two tasks used the same or different response modalities. Together, these two results provide substantial support for the idea that optimality considerations influence the use of serial versus parallel processing modes in PRP tasks.

*Can the Bottleneck Model Account for the Results?*

Although the present results are quite consistent with—and indeed were predicted from—the hypothesis that parallel processing can be used in the PRP task when it is optimal, it is important to examine carefully the question of how the bottleneck model might be modified to account for these results. As noted in the introduction, the modal form of that model predicts that RT2 would depend on the actual SOA but not on the relative likelihoods of the various SOAs (cf. Equations 2 and 4 of Pashler & Johnston, 1989). Perhaps, however, a rather straightforward modification of the bottleneck model could also account for the effect of the likely SOA as well as the actual SOA.

We cannot find any such modification. The bottleneck model is usually elaborated in one of two ways to account for discrepant results (cf. Navon & Miller, 2002). One possible elaboration is to suppose that participants sometimes group their responses in the two tasks, thus emitting both responses at nearly the same time (e.g., Borger, 1963; Pashler & Johnston, 1989). In the present experiments, it seems plausible that participants would have a stronger tendency to group responses when a short SOA was likely than when a long SOA was likely (e.g., Ivry, et al., 1998; Lien, et al., 2003; Pashler & Johnston, 1989; Ruthruff, Pashler, & Hazeltine, 2003). Could such a change in the frequency of grouping explain the results?

Without a detailed model of response grouping, it is not clear exactly what changes in RT1 and RT2 this elaboration of the bottleneck model would predict. Therefore, it is impossible to prove that no model of this sort could give a good account of our results. Four aspects of the results suggest, however, that response grouping was not responsible for the effects of SOA distribution observed in these experiments. First, median-split analyses provided evidence against the idea that the effects were mediated by grouping. For all three experiments, we used median splits to divide trials into those with relatively small versus relatively large IRIs. In all experiments, the effects of SOA distribution were statistically as large or larger in the trials with

relatively long IRIs as in the trials with relatively short ones. Based on the standard assumption that response grouping tends to produce short IRIs, this implies that these SOA distribution effects were not mediated by grouping. Second, the effects of the SOA distribution were rather insensitive to the cutoffs used to exclude grouped responses. The presented results were computed with a cutoff of 100 ms, but almost identical results were also obtained using cutoffs of 50 ms and 140 ms. Surely the number of grouped-response trials included in the analyses should depend on the cutoff. Therefore, effects that were due to grouping should have changed as a function of the cutoff size, but there were no such changes. Third, the sizes of the SOA distribution effects were not well correlated across experiments with the percentages of trials having short IRIs. Comparison of Figures 3, 5, and 7 shows that the percentages of trials with IRIs less than 100 ms decreased substantially across the three experiments—probably partly due to more instructional emphasis on avoiding grouping. In contrast, the effects of SOA distribution tended to increase across experiments, clearly suggesting that these effects were not a result of grouping. Fourth, although Experiment 1 yielded a substantially larger percentage of trials with IRI$<$ 100 ms in condition SF than in condition LF—as would be needed to account for the SOA distribution effect in terms of grouping—the corresponding differences were rather small in Experiments 2 and 3. Obviously, if there was little or no difference in grouping for SF versus LF, then it is extremely unlikely that grouping is responsible for other differences between these conditions.

Finally, as noted earlier, there is also a strong theoretical argument against the claim that response grouping is responsible for the effects of SOA distribution. In brief, grouping models assume that the Task 1 response is held back until the Task 2 response is ready (e.g., Borger, 1963; Pashler & Johnston, 1989). They can easily explain the increase of RT1 in the SF condition by assuming that R1 is held back more often in this condition. These models have difficulty explaining any effects on RT2, however, because Task 2 is processed normally while R1 is waiting. Regardless of how often Task 1 responses are held back, the function relating RT2 to SOA should still have a slope of -1. It would therefore be extremely difficult for a grouping model to explain the observed effects of SOA distribution on the RT2 slope.

The other standard modification to bottleneck models uses the concept of task preparation. For example, the RT for a given task is usually larger when that task is the first task in the PRP

paradigm than when the same task is performed in isolation (e.g., Pashler & Johnston, 1989). Although this change in RT is not predicted by bottleneck models, it can be explained by arguing that participants prepare better for a task performed in isolation than for the same task performed in the PRP situation. As another example, increases in RT1 associated with greater Task 2 difficulty can be reconciled with a bottleneck model by arguing that participants reduce their preparation for Task 1 when Task 2 is especially difficult. In addition, Gottsdanker (1979) found that RT2 was lengthened in the PRP paradigm even when the first stimulus was not presented. Given that this RT2 increase could not be attributed to any delay associated with processing S1, this result also provides further evidence that task preparation is important. In general, the concept of preparation could be extended to explain a variety of differences between any kind of blocked conditions, because the level of preparation for a given task could change across blocks.

To explain the present results in terms of task preparation, it would be reasonable to assume that participants prepared almost completely for Task 1 when SOA is usually long. When SOA is usually short, however, they might prepare more equally for the two tasks. Decreasing preparation for Task 1 when a short SOA was likely would increase RT1 in that condition, consistent with the results. The preparation account fails, however, to explain the robust interaction between the actual SOA and the likely SOA with respect to RT2. To substantiate this claim, Appendix D presents a specific model incorporating differential preparation and shows that this model is not consistent with the actual form of the interactions observed in these experiments.

Finally, another possibility to account for the present results in terms of the bottleneck model is to attribute the effects of SOA distribution on RT2 to changes in temporal preparation.[8] It is well-known that participants' uncertainty about when a stimulus will appear affects RT (e.g., Niemi & Näätänen, 1981). When short SOAs are likely, participants might be especially prepared for Task 2 at short SOA values. In contrast, when long SOAs are likely, temporal preparation for Task 2 might be more pronounced at these long SOA values. Thus, the SOA effect would be modulated by the SOA distribution condition, producing an interaction of SOA and SOA distribution. This predicted interaction, however, is only consistent with the present RT2 results if temporal preparation shortens the duration of post-bottleneck processes in Task 2 (see Appendix D). Recent studies of temporal preparation do not support this view of temporal

preparation effects (Hackley & Valle-Inclán, 2003; Müller-Gethmann, Ulrich & Rinkenauer, 2003). These studies suggest that temporal preparation operates on processes before or during the bottleneck rather than after it. In addition, temporal preparation cannot directly account for the effect of SOA distribution on RT1, because this manipulation does not affect temporal uncertainty regarding the onset of S1.

*Implications for Dual-Task Performance*

Two rather general conclusions about dual-task processing are supported by the present evidence that the likelihood of long versus short SOAs influences the use of serial versus parallel processing modes. The first and more important conclusion is that the results cast doubt on the idea of a structural bottleneck limiting dual-task performance by requiring strictly serial processing. At a metatheoretical level, our analysis provides a clearcut alternative explanation for the finding that processing generally appears to be serial in the PRP paradigm. Specifically, the optimization framework suggests that this mode of processing would almost always be more efficient than parallel processing. Assuming that the mode of performance tends to be adjusted, consciously or unconsciously, toward the optimal mode, then, we would normally expect participants to process in a serial mode even if there were no structural bottleneck preventing parallel processing. Furthermore, at an empirical level, the effects of SOA distribution provide direct evidence that a shift toward more parallel processing is possible when it is more efficient than serial processing, at least under some conditions. Given that such a shift is incompatible with structural bottleneck models, evidence of it clearly tends to support the conjecture that performance optimization—not a structural limitation—is the reason why processing is usually serial in the PRP paradigm.

Our first conclusion is perfectly in keeping with previous claims that certain constraints built into typical PRP tasks tend to make serial processing more efficient than parallel processing (cf. Meyer & Kieras, 1997a, 1997b; Navon & Miller, 2002; Tombu & Jolicœur, 2003). Indeed, our optimization framework identifies a further constraint, not previously developed formally, working to encourage serial processing. Serial processing has previously been shown to be more efficient than parallel processing within the context of specific models (e.g., Logan & Gordon, 2001), and

the present metatheoretical analysis extends this conclusion to broad classes of models.

A caveat is in order, however, regarding the extent of parallel processing. Although the present results provide evidence that parallel processing is possible, the observed data patterns were still consistent with a processing mode that was primarily serial. For example, the function relating RT2 to SOA had quite a steep slope in Experiments 1 and 2—nearly minus 1, approaching the prediction of the bottleneck model—even in the conditions encouraging parallel processing. This finding is consistent with the claim that our experimental designs were only partly successful in encouraging such processing, so it occurred in only a small percentage of participants or trials, although as noted earlier slopes close to -1 are also compatible with certain parallel models. Even if participants did process primarily in a serial model, however, this may have been partly because they failed to optimize fully their performance in response to changing task parameters, as had also been observed in other situations (e.g. Dickman & Meyer, 1988). Alternatively, it is also possible that the manipulation of SOA distribution was simply not powerful enough to create a condition in which parallel processing was always optimal. The optimization framework (e.g., Equation 3) shows that the optimal processing mode is also influenced by the ratio of processing times in the two modes, $X_p/X_s$, and in our tasks this ratio may simply have been so large that serial processing was always optimal.

Fortunately, the optimal processing framework that led us to manipulate the likelihood of small versus large SOAs also suggests other manipulations that could be used in addition to encourage parallel processing even further. For example, any manipulation that speeded parallel processing relative to serial processing (i.e., that made $X_p$ not too much slower than $X_s$) would also encourage participants to handle the two tasks in parallel. Likewise, using a harder Task 1 and an easier Task 2 would tend to encourage parallel processing (e.g., see Appendix B). Obviously, if the present optimization framework is correct, then combining all of these manipulations should produce a level of parallel processing that should be easier to discriminate from the purely serial mode. Thus, another strength of the framework is that it provides experimenters with specific guidance about how to create conditions that would be optimal for parallel processing.

The second general conclusion is that our data emphasize the importance of relative

preparation for task 1 versus task 2 in PRP paradigms (cf. De Jong, 1995; Gottsdanker, 1979, 1980; Luria & Meiran, 2003). All three experiments show that the effects of the actual SOA are modulated by the SOA distribution, and this implicates influences of expectancy and advance preparation on processing. It appears that the relative emphasis on serial versus parallel processing can be adjusted to some extent in advance of the trial, based on the anticipated efficiency of each mode. Moreover, an effect of advance preparation suggests that the system cannot make instantaneous adjustments to optimize its processing mode. If it could, no advance preparation would be needed. Within a capacity model, for example, if the reallocation of capacity to tasks could be accomplished instantaneously within a trial, then the advance expectation of a short versus long SOA should have no effect on RT2, at least when the actual SOA is long. Because expectations do have an effect at long SOAs, it appears that participants are not capable of readjusting their processing mode instantaneously during a trial as the passage of time makes it evident that the current trial's SOA is long rather than short.

*Methodological Implications of the Optimization Framework*

It is also informative to reevaluate previous attempts to test bottleneck models in the light of the present optimization framework and experimental results. If serial processing is generally optimal, of course, then it is not surprising that previous investigators would mostly have found evidence for serial processing, even if parallel processing were possible. Thus, the optimization framework undermines the position that the ubiquity of serial processing is most likely caused by a structural bottleneck.

The present results also have specific implications for a number of studies in which researchers sought to test bottleneck models by setting up conditions with strong incentives to encourage participants to use parallel processing if it were possible. Although some did indeed find evidence for parallel processing (e.g., Hazeltine, et al., 2002; Schumacher, Lauber, Glass, Zurbriggen, Gmeindl, Kieras, & Meyer, 1999), many found evidence that processing was still serial and concluded that parallel processing was therefore likely to be impossible under conditions similar to those used in their experiments. Reconsideration in terms of the optimization framework, however, raises doubts about whether these investigators truly succeeded

in setting up optimal conditions for parallel processing.

For example, Pashler (1994b) used a task in which S1 and S2 were sometimes presented simultaneously, and he instructed participants to respond as quickly as possible, emphasizing both tasks equally. He claimed that "if people can split their capacity among the tasks arbitrarily, our instructions would seem to encourage a more or less even split about as effectively as could be done" (p. 335). The present optimization framework appears to contradict this claim, however. With equal task emphasis, it seems especially plausible that participants would try to minimize TRT = RT1 + RT2. Our results indicate that serial processing is more likely to do that than is parallel processing, especially given that relative long SOAs (±500ms and ±1000 ms) were used in 80% of the trials in his experiment. Similarly, Ruthruff, Pashler, and Hazeltine (2003) also attempted to find evidence for parallel processing with equal task emphasis, extending the work of Pashler (1994b) by using tasks with different response modalities to eliminate any possible response initiation bottleneck. Their first experiment also included many long SOAs, however, and their second included many single-task trials (effectively $SOA = \infty$). Based on the present analysis, then, it again seems rather doubtful that parallel processing would have been optimal. If Pashler (1994b) and Ruthruff et al. (2003) did not actually optimize their tasks for parallel processing, of course, then the fact that they found evidence of serial processing still cannot be taken as evidence that parallel processing was prevented by a structural bottleneck.[9]

In addition to these two studies, the literature on the PRP paradigm includes quite a few other examples of experiments in which investigators have tried to encourage parallel processing, especially by providing participants with (a) strong incentives for fast and accurate performance, and (b) extensive practice at the tasks (e.g., Spelke, Hirst, & Neisser, 1976; Van Selst, Ruthruff, & Johnston, 1999). It is beyond the scope of this article to review all of these experiments, but two general points can be made. First, strong incentives in and of themselves need not encourage parallel processing and may instead discourage it. We imagine that such incentives make participants work even harder to optimize their performance, but the present analysis shows that under many conditions this is more likely to force them into a serial mode than into a parallel one. The use of strong incentives to encourage parallel processing in past studies seems to have been based on an implicit assumption that if people would just work harder, they could do two

things at once with little or no interference. This assumption is just as inconsistent with limited-capacity models as it is with the bottleneck model, however, so it seems inappropriate to rely on it when trying to find evidence for parallel processing.

Second, it is not necessarily true that extensive practice would make parallel processing more optimal than serial processing. Within the context of the present optimization framework, it is natural to suppose that practice decreases the time needed to perform a task. In terms of the framework, this would surely produce reductions in $X_s$ and $X_p$, but it is not at all obvious to us why it should differentially favor $X_p$ to the point of satisfying the inequality $X_p < 1.5 \cdot X_s$. If a given amount of practice reduces both $X_s$ and $X_p$ by the same proportion, for example, then no amount of practice will reverse the inequality. Moreover, if participants process serially early in practice, then they gain practice at serial processing, not parallel processing, which would presumably tend to reduce $X_s$ more than $X_p$, assuming that the greatest benefits accrue to the mode actually being practiced.

Finally, in retrospect it seems remarkable that the issue of the optimality of serial versus parallel processing modes has previously been largely overlooked in the literature on the PRP paradigm (but for an exception, see Wickens & Seidler, 1997). Given that people tend to perform optimally in many circumstances (e.g., Anderson, 1990; Ernst & Banks, 2002; Geisler & Diehl, 2003; Legge, Hooven, Klitz, Mansfield, & Tjan, 2002; Meyer, Abrams, Kornblum, Wright, & Smith, 1988; Movellan & McClelland, 2001; Navon, 1978; Shaw & Shaw, 1977; Sperling, 1984), it is reasonable to expect them to behave as optimally as possible in this paradigm as well. Without knowing whether serial or parallel processing is optimal in a given situation, however, investigators can hardly be confident in concluding that the unused processing mode is impossible. We hope that future studies will extend the present analysis of optimality to identify and study situations in which participants actually should attempt to process in parallel if they can do so. Only serial processing observed in such situations would be convincing evidence that parallel processing is prevented by a structural bottleneck.

References

Allport, D. A. (1980). Attention and performance. In G. Claxton (Ed.), *Cognitive psychology - New directions* (pp. 112–153). London: Routledge & Kegan Paul.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Borger, R. (1963). The refractory period and serial choice-reactions. *Quarterly Journal of Experimental Psychology, 15,* 1–12.

Conway, R. W., Maxwell, W. L., & Miller, L. W. (1967). *Theory of scheduling.* Reading, MA: Addison-Wesley.

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). The role of preparation in overlapping-task performance. *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology, 48,* 2–25.

Dickman, S. J., & Meyer, D. E. (1988). Impulsivity and speed-accuracy tradeoffs in information processing. *Journal of Personality & Social Psychology, 54,* 274–290.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429–433.

Geisler, W. S., & Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science, 27,* 379–402.

Ghez, C. (1991). Voluntary movement. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (Eds.), *Principles of neural science (3rd ed.)* (pp. 609–625). Norwalk, Conn.: Appleton & Lange.

Gottsdanker, R. (1979). A psychological refractory period or an unprepared period? *Journal of Experimental Psychology: Human Perception and Performance, 5,* 208–215.

Gottsdanker, R. (1980). The ubiquitous role of preparation. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 355–372). Amsterdam: North Holland.

Hackley, S. A., & Valle-Inclán, F. (2003). Which stages of processing are speeded by a warning signal? *Biological Psychology, 64,* 27–45.

Hazeltine, E., Teague, D., & Ivry, R. B. (2002). Simultaneous dual-task performance reveals parallel response selection after practice. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 527–545.

Hirst, W., Spelke, E. S., Reaves, C. C., Caharack, G., & Neisser, U. (1980). Dividing attention without alternation or automaticity. *Journal of Experimental Psychology: General, 109,* 98–117.

Hommel, B. (1998). Automatic stimulus-response translation in dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1368–1384.

Hubel, D. H. (1979). The brain. *Scientific American, 241,* 39–47.

Ivry, R. B., Franz, E. A., Kingstone, A., & Johnston, J. C. (1998). The psychological refractory period effect following callosotomy: Uncoupling of lateralized response codes. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 463–480.

Jersild, A. T. (1927). Mental set and shift. *Archives of Psychology*, no. 89.

Jolicœur, P., & Dell'Acqua, R. (1999). Attentional and structural constraints on visual encoding. *Psychological Research, 62,* 154–164.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Kandel, E. R. (1991). Perception of motion, depth, and form. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (Eds.), *Principles of neural science (3rd ed.)* (pp. 440–466). Norwalk, Conn.: Appleton & Lange.

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

Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., & Tjan, B. S. (2002). Mr. Chips 2002: New insights from an ideal-observer model of reading. *Vision Research, 42,* 2219–2234.

Lien, M. C., Schweickert, R., & Proctor, R. W. (2003). Task switching and response correspondence in the psychological refractory period paradigm. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 692–712.

Logan, G. D., & Delheimer, J. A. (2001). Parallel memory retrieval in dual-task situations: II. Episodic memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27,* 668–685.

Logan, G. D., & Gordon, R. D. (2001). Executive control of visual attention in dual-task situations. *Psychological Review, 108,* 393–434.

Logan, G. D., & Schulkind, M. D. (2000). Parallel memory retrieval in dual-task situations: I. Semantic memory. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1072–1090.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* Oxford: Oxford University Press.

Luria, R., & Meiran, N. (2003). Online order control in the psychological refractory period paradigm. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 556–574.

Martin, J. H. (1991). Coding and processing of sensory information. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (Eds.), *Principles of neural science (3rd ed.)* (pp. 329–340). Norwalk, Conn.: Appleton & Lange.

McLeod, P. (1977). A dual task response modality effect: Support for multiprocessor models of attention. *Quarterly Journal of Experimental Psychology, 29,* 651–667.

McLeod, P. (1978). Does probe RT measure central processing demand? *Quarterly Journal of Experimental Psychology, 30,* 83–89.

McLeod, P. (1980). What can probe RT tell us about the attention demands of movement? In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 579–590). Amsterdam: North Holland.

Meyer, D. E., Abrams, R. A., Kornblum, S., Wright, C. E., & Smith, J. E. K. (1988). Optimality in human motor performance: Ideal control of rapid aimed movements. *Psychological Review, 95,* 340–370.

Meyer, D. E., & Kieras, D. E. (1997a). A computational theory of executive cognitive processes and multiple-task performance: I. Basic mechanisms. *Psychological Review, 104,* 3–65.

Meyer, D. E., & Kieras, D. E. (1997b). A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological Review, 104,* 749–791.

Meyer, D. E., & Kieras, D. E. (1999). Precis to a practical unified theory of cognition and action: Some lessons from EPIC computational models of human multiple-task performance. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 17–88). Cambridge, MA: MIT Press.

Miller, J. O., & Ulrich, R. (2003). Simple reaction time and statistical facilitation: A parallel grains model. *Cognitive Psychology, 46,* 101–151.

Movellan, J. R., & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review, 108,* 113–148.

Müller-Gethmann, H., Ulrich, R., & Rinkenauer, G. (2003). Locus of the effect of temporal preparation: Evidence from the lateralized readiness potential. *Psychophysiology, 40,* 597–611.

Navon, D. (1978). The importance of being conservative: Some reflections on human Bayesian behaviour. *British Journal of Mathematical & Statistical Psychology, 31,* 33–48.

Navon, D. (1984). Resources — A theoretical soup stone? *Psychological Review, 91,* 216–234.

Navon, D., & Gopher, D. (1979). On the economy of the human information processing system. *Psychological Review, 86,* 214–255.

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

Navon, D., & Miller, J. O. (2002). Queuing or sharing? A critical evaluation of the single-bottleneck notion. *Cognitive Psychology, 44,* 193–251.

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological Bulletin, 89,* 133–162.

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

Pashler, H. E. (1990). Do response modality effects support multiprocessor models of divided attention? *Journal of Experimental Psychology: Human Perception and Performance, 16,* 826–842.

Pashler, H. E. (1994a). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116,* 220–244.

Pashler, H. E. (1994b). Graded capacity-sharing in dual-task interference? *Journal of Experimental Psychology: Human Perception and Performance, 20,* 330–342.

Pashler, H. E., & Johnston, J. C. (1989). Chronometric evidence for central postponement in temporally overlapping tasks. *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology, 41,* 19–45.

Rauschecker, J. P. (1998). Parallel processing in the auditory cortex of primates. *Audiology & Neuro-Otology, 3,* 86–103.

Rinkenauer, G., Ulrich, R., & Wing, A. M. (2001). Brief bimanual force pulses: Correlations between the hands in force and time. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 1485–1497.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations.* Cambridge, MA: MIT Press.

Ruthruff, E. D., Johnston, J. C., & Van Selst, M. (2001). Why practice reduces dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 3–21.

Ruthruff, E. D., Johnston, J. C., Van Selst, M., Whitsell, S., & Remington, R. (2003). Vanishing dual-task interference after practice: Has the bottleneck been eliminated or is it merely latent? *Journal of Experimental Psychology: Human Perception and Performance, 29,* 280–289.

Ruthruff, E. D., Pashler, H. E., & Hazeltine, E. (2003). Dual-task interference with equal task emphasis: Graded capacity sharing or central postponement? *Perception & Psychophysics, 65,* 801–816.

Ruthruff, E. D., Pashler, H. E., & Klaassen, A. (2001). Processing bottlenecks in dual-task performance: Structural limitation or strategic postponement? *Psychonomic Bulletin & Review, 8,* 73–80.

Schumacher, E. H., Lauber, E. J., Glass, J. M., Zurbriggen, E. L., Gmeindl, L., Kieras, D. E., & Meyer, D. E. (1999). Concurrent response-selection processes in dual-task performance: Evidence for adaptive executive control of task scheduling. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 791–814.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001). Virtually perfect time sharing in dual-task performance: Uncorking the central cognitive bottleneck. *Psychological Science, 12,* 101–108.

Schwarz, W., & Ischebeck, A. (2001). On the interpretation of response time vs onset asynchrony functions: Application to dual-task and precue-utilization paradigms. *Journal of Mathematical Psychology, 45,* 452–479.

Schweickert, R., & Boggs, G. J. (1984). Models of central capacity and concurrency. *Journal of Mathematical Psychology, 28,* 223–281.

Smith, M. C. (1969). The effect of varying information on the psychological refractory period. *Acta Psychologica, 30,* 220–231.

Spector, A., & Biederman, I. (1976). Mental set and mental shift revisited. *American Journal of Psychology, 89,* 669–679.

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

Sperling, G. (1984). A unified theory of attention and signal detection. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 103–182). New York: Academic Press.

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

Tombu, M., & Jolicœur, P. (2002). All-or-none bottleneck versus capacity sharing accounts of the psychological refractory period phenomenon. *Psychological Research, 66,* 274–286.

Tombu, M., & Jolicœur, P. (2003). A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 3–18.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes.* Cambridge: Cambridge University Press.

Van Selst, M., Ruthruff, E., & Johnston, J. C. (1999). Can practice eliminate the Psychological Refractory Period effect? *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1268–1283.

Wässle, H., Grunert, U., Rohrenbeck, J., & Boygott, B. B. (1990). Retinal ganglion cell density and cortical magnification factor in the primate. *Vision Research, 30,* 1897–1911.

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. (1959). Evidence of a single-channel decision mechanism limiting performance in a serial reaction task. *Quarterly Journal of Experimental Psychology, 11,* 193–210.

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

Wickens, C. D. (1976). The effects of divided attention on information processing in manual tracking. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 1–13.

Wickens, C. D., & Seidler, K. S. (1997). Information access in a dual-task context: Testing a model of optimal strategy selection. *Journal of Experimental Psychology:Applied, 3,* 196–215.

Wühr, P., & Müsseler, J. (2002). Blindness to response-compatible stimuli in the psychological refractory period paradigm. *Visual Cognition, 9,* 421–457.

## Appendix A

## Extension of Optimization Framework to Three-Stage Models With Random Stage Durations

This appendix extends the optimization framework presented in the main text to a more complicated situation that corresponds more closely to canonical RT models considered within the literature on the PRP paradigm. Specifically, we consider here the case in which (a) each task is carried out by a sequence of three stages corresponding to pre-bottleneck perceptual processes, central bottleneck processes, and post-bottleneck motor processes (cf. Navon & Miller, 2002; Pashler, 1984; Pashler & Johnston, 1989; Schwarz & Ischebeck, 2001; Tombu & Jolicœur, 2003), and (b) the duration of each stage is a random variable rather than a constant. We show that virtually the same conclusions reached in the main text from the simpler model can also be reached from this more complex model.

Figure A1 depicts the model underlying this analysis. Processing with the serial mode is shown in the top half of the figure, and processing with the parallel mode is shown in the bottom half. We consider formally only the limiting case of SOA = 0. Note that the use of the serial versus parallel processing mode affects only the operation of the central stages for the two tasks. The perceptual and motor stages of the two tasks can occur in parallel without dual-task interference. The durations of the central stages are, however, shorter when these stages are carried out serially than when they are carried out in parallel.

---

Insert Figure A1 about here

---

With serial processing, the RTs for the two tasks are

$$\mathrm{RT1}_s = \mathrm{A1} + \mathrm{B1}_s + \mathrm{C1} \tag{5}$$

$$\mathrm{RT2}_s = \mathrm{A1} + \mathrm{B1}_s + \mathrm{B2}_s + \mathrm{C2} \tag{6}$$

Note that we are assuming that Task 1 is always processed first when processing occurs serially. The appropriateness of this assumption—at least to a close approximation—is strongly supported

by the fact that R1 is almost always emitted before R2 in tasks where there is a strong expectation that S1 will be presented before S2 (De Jong, 1995), as is true in almost all PRP experiments. To further insure the accuracy of this assumption with respect to our own data, we excluded the rare trials in which R2 was emitted before R1.

In contrast, with parallel processing, the predicted RTs are

$$\mathrm{RT1}_p = \mathrm{A1} + \mathrm{B1}_p + \mathrm{C1} \tag{7}$$

$$\mathrm{RT2}_p = \mathrm{A2} + \mathrm{B2}_p + \mathrm{C2} \tag{8}$$

As in the main text, we are assuming at this point that the time needed for central parallel processing is independent of whether two processes need the central process at the same time. For example, $\mathrm{B1}_p$ is not reduced in trials where Task 1 has a big head start in central processing (i.e., when $\mathrm{A1} << \mathrm{A2}$). If capacity could be reallocated quickly within the trial, parallel processing might be faster when the tasks arrived at the central processor at rather different times, contrary to this assumption. The more complicated case in which parallel processing times depend on temporal overlap of demand is considered in Appendix C.

To determine whether the serial or parallel mode minimizes the overall task time, it is sufficient to examine

$$\Delta = \mathrm{E}\left[\mathrm{RT1}_p + \mathrm{RT2}_p\right] - \mathrm{E}\left[\mathrm{RT1}_s + \mathrm{RT2}_s\right] \tag{9}$$

The serial mode should be preferred when $\Delta$ is positive; the parallel mode, when $\Delta$ is negative. Inserting the values from Equations 5–8 and simplifying yields

$$\Delta = \mathrm{E}\left[\mathrm{A2}\right] - \mathrm{E}\left[\mathrm{A1}\right] + \mathrm{E}\left[\mathrm{B1}_p\right] + \mathrm{E}\left[\mathrm{B2}_p\right] - 2\mathrm{E}\left[\mathrm{B1}_s\right] - \mathrm{E}\left[\mathrm{B2}_s\right] \tag{10}$$

To see how this quantity depends on the durations of the central processes, it is convenient to assume that $\mathrm{E}\left[\mathrm{A1}\right] = \mathrm{E}\left[\mathrm{A2}\right]$ and that parallel processing time is proportional to serial processing time (i.e., $\mathrm{Bi}_p = \alpha \cdot \mathrm{Bi}_s$ for some $\alpha > 1$). In that case we obtain

$$\Delta = (\alpha - 2) \cdot \mathrm{E}\left[\mathrm{B1}_s\right] + (\alpha - 1) \cdot \mathrm{E}\left[\mathrm{B2}_s\right] \tag{11}$$

This quantity clearly increases with $\alpha$, so there is a clear advantage for serial processing if the time needed for parallel processing is sufficiently longer than that needed for serial processing

(e.g., $\alpha > 2$). Note also that the parallel mode tends to gain more when the second task is hard, that is, when $\mathrm{E}[\mathrm{B2}_s] > \mathrm{E}[\mathrm{B1}_s]$. For the special case of $\mathrm{E}[\mathrm{B1}_s] = \mathrm{E}[\mathrm{B2}_s]$, which is analogous to the situation of equal task difficulty considered in the main text, the crossover point again occurs at $\alpha = 1.5$. With values of $\alpha$ greater than that, serial processing is better, but with values of $\alpha$ less than that, parallel processing is better.

As in the main text, then, serial processing may be more efficient than parallel processing even when SOA = 0, if the time needed for parallel processing in the central stage is sufficiently longer than the time needed for serial processing in that same stage. Furthermore, it is easy to see from Figure A1 that the advantage for serial processing grows as SOA increases. For the serial mode, increasing SOA decreases $\mathrm{RT2}_s$—and concomitantly decreases the overall task time—until the waiting time is eliminated, and then $\mathrm{RT2}_s$ asymptotes at its minimal value of $\mathrm{A2} + \mathrm{B2}_s + \mathrm{C2}$. For the parallel mode, however, increasing SOA has no effect on either $\mathrm{RT1}_p$ or $\mathrm{RT2}_p$, and hence produces no decrease in the overall task time.

## Appendix B

## Optimal Scheduling in a Simple Capacity Model

This appendix considers the special case of standard capacity models (e.g., Navon & Gopher, 1979; Navon & Miller, 2002; Tombu & Jolicœur, 2003) in terms of the optimization framework developed more generally in the main text. Specifically, we start from the premises that (a) each task $i = 1, 2$ can be represented in terms of the total amount of work needed for its completion, $W_i$; (b) the information processing system has a fixed total capacity, $\mathcal{N}$, that can be divided between tasks; (c) if the allocations of capacity to Tasks 1 and 2 are $\mathcal{N} \cdot z$ and $\mathcal{N} \cdot (1 - z)$, $0 < z < 1$, respectively, then the times needed for completion of the tasks are $\frac{W_1}{\mathcal{N} \cdot z}$ and $\frac{W_2}{\mathcal{N} \cdot (1-z)}$, respectively.

*Case 1: SOA = 0*

We first determine the optimal allocation of capacity to the two tasks. In other words, what value of $z$ minimizes the total time under the parallel mode? In this mode the total time is

$$\mathrm{TRT}_p = \frac{W_1}{\mathcal{N} \cdot z} + \frac{W_2}{\mathcal{N} \cdot (1 - z)} \tag{12}$$

To find the minimum of $\mathrm{TRT}_p$, we compute its derivative with respect to $z$, which is

$$\frac{d\mathrm{TRT}_p}{dz} = \frac{-W_1}{\mathcal{N} z^2} + \frac{W_2}{\mathcal{N} \cdot (1 - z)^2} \tag{13}$$

Setting this derivative equal to 0 and solving for $z$ yields

$$z = \frac{W_1 - \sqrt{W_1 \cdot W_2}}{W_1 - W_2} \tag{14}$$

which can be further simplified as follows:

$$\begin{aligned} z &= \frac{W_1 - \sqrt{W_1 \cdot W_2}}{W_1 - W_2} && (15)\\ &= \frac{\sqrt{W_1} \cdot \sqrt{W_1} - \sqrt{W_1} \cdot \sqrt{W_2}}{(\sqrt{W_1} - \sqrt{W_2}) \cdot (\sqrt{W_1} + \sqrt{W_2})} && (16)\\ &= \frac{\sqrt{W_1} \cdot (\sqrt{W_1} - \sqrt{W_2})}{(\sqrt{W_1} - \sqrt{W_2}) \cdot (\sqrt{W_1} + \sqrt{W_2})} && (17)\\ &= \frac{\sqrt{W_1}}{\sqrt{W_1} + \sqrt{W_2}} && (18) \end{aligned}$$

Thus, the optimal allocation of capacity—to minimize total processing time—is to assign proportions $z = \frac{\sqrt{W_1}}{\sqrt{W_1}+\sqrt{W_2}}$ and $1 - z = \frac{\sqrt{W_2}}{\sqrt{W_1}+\sqrt{W_2}}$ to Tasks 1 and 2, respectively. With this allocation, the total time needed for parallel processing is

$$\begin{aligned} \mathrm{TRT}_{p,min} &= \frac{W_1}{\mathcal{N} \cdot \frac{\sqrt{W_1}}{\sqrt{W_1}+\sqrt{W_2}}} + \frac{W_2}{\mathcal{N} \cdot \frac{\sqrt{W_2}}{\sqrt{W_1}+\sqrt{W_2}}} && (19) \\ &= \frac{\sqrt{W_1}+\sqrt{W_2}}{\mathcal{N}} \cdot \left(\sqrt{W_1}+\sqrt{W_2}\right) && (20) \\ &= \frac{1}{\mathcal{N}} \cdot \left(W_1 + W_2 + 2 \cdot \sqrt{W_1 \cdot W_2}\right) && (21) \end{aligned}$$

In contrast, the total time for serial processing is

$$\begin{aligned} \mathrm{TRT}_s &= 2 \cdot \frac{W_1}{\mathcal{N}} + \frac{W_2}{\mathcal{N}} && (22) \\ &= \frac{1}{\mathcal{N}} \cdot (2 \cdot W_1 + W_2) && (23) \end{aligned}$$

Parallel processing is better than serial processing when

$$\begin{aligned} \mathrm{TRT}_s &> \mathrm{TRT}_{p,min} && (24) \\ \frac{1}{\mathcal{N}} \cdot (2 \cdot W_1 + W_2) &> \frac{1}{\mathcal{N}} \cdot \left(W_1 + W_2 + 2 \cdot \sqrt{W_1 \cdot W_2}\right) && (25) \\ W_1 &> 2 \cdot \sqrt{W_1 \cdot W_2} && (26) \\ \frac{1}{2} &> \frac{\sqrt{W_1 \cdot W_2}}{\sqrt{W_1 \cdot W_1}} && (27) \\ \frac{1}{2} &> \frac{\sqrt{W_2}}{\sqrt{W_1}} && (28) \\ \frac{1}{4} &> \frac{W_2}{W_1}, && (29) \end{aligned}$$

or when $W_1 > 4 \cdot W_2$, that is, if the amount of work for Task 1 is at least four times as large as the amount of work for Task 2.

To make these ideas more concrete, Table B1 illustrates two numerical examples.

---

Insert Table B1 about here

---

*Case 2: SOA* $> 0$

Conditions with SOA $> 0$ tend to be more favorable for serial processing, relative to parallel processing, than conditions with SOA $= 0$. That is, if serial processing is more efficient

than parallel processing at SOA = 0, then its advantage is even larger with SOA > 0. Alternatively, if serial processing is less efficient than parallel processing at SOA = 0, its disadvantage decreases and may even reverse with SOA > 0. This is because, relative to the case of simultaneous onset, the time needed for serial processing decreases and the time needed for parallel processing remains unchanged.

*Case 3: Extension to Three-Stage Models*

In most information processing models the capacity-limited central stage is preceded by a perceptual stage and followed by a motor stage, and both of these additional stages are assumed to be capable of operating in an unlimited-capacity parallel fashion if there are no structural limits between tasks. How would the conclusions of the present capacity-based analysis differ if the durations of these additional unlimited-capacity stages were taken into account?

Suppose that Tasks 1 and 2 begin with perceptual stages having durations A1 and A2, respectively, and that they finish with motor stages having durations $C1$ and $C2$, and let the total processing time for this three-stage model be denoted as TRT′. In the parallel processing mode, the total time is

$$\begin{aligned} \mathrm{TRT}'_p &= \left(\mathrm{A1} + \frac{W_1}{\mathcal{N} \cdot z} + C1\right) + \left(\mathrm{A2} + \frac{W_2}{\mathcal{N} \cdot (1-z)} + C2\right) && (30) \\ &= \mathrm{A1} + \mathrm{A2} + C1 + C2 + \mathrm{TRT}_p && (31) \end{aligned}$$

$\mathrm{TRT}'_p$ has the same minimum with respect to $z$ as does $\mathrm{TRT}_p$, because the perceptual and motor stage durations are additive constants. Thus, the total time needed with the optimal allocation mode is

$$\begin{aligned} \mathrm{TRT}'_{p,min} &= \mathrm{A1} + \frac{W_1}{\mathcal{N} \cdot \frac{W_1}{\sqrt{W_1}+\sqrt{W_2}}} + C1 + \mathrm{A2} + \frac{W_2}{\mathcal{N} \cdot \frac{W_2}{\sqrt{W_1}+\sqrt{W_2}}} + C2 && (32) \\ &= \mathrm{A1} + \mathrm{A2} + C1 + C2 + \mathrm{TRT}_{p,min} && (33) \end{aligned}$$

Assuming that the motor processes can operate in parallel with other processes without interference and that the perceptual processing for Task 2 can be finished while Task 1 is being processed (i.e., $\mathrm{A1} + \frac{W_1}{\mathcal{N}} \geq \mathrm{A2}$), the total time needed for serial processing is

$$\begin{aligned} \mathrm{TRT}'_s &= 2 \cdot \left(\mathrm{A1} + \frac{W_1}{\mathcal{N}}\right) + C1 + \frac{W_2}{\mathcal{N}} + C2 && (34) \\ &= 2 \cdot \mathrm{A1} + C1 + C2 + \mathrm{TRT}_s && (35) \end{aligned}$$

With preliminary perceptual stages, then, parallel processing is better than serial processing when

$$\begin{aligned}
\mathrm{TRT}'_s &> \mathrm{TRT}'_p && (36)\\
2 \cdot \mathrm{A1} + C1 + C2 + \mathrm{TRT}_s &> \mathrm{A1} + \mathrm{A2} + C1 + C2 + \mathrm{TRT}_p && (37)\\
\mathrm{A1} + \mathrm{TRT}_s &> \mathrm{A2} + \mathrm{TRT}_p && (38)\\
\mathrm{A1} + \frac{1}{\mathcal{N}} \cdot (2 \cdot W_1 + W_2) &> \mathrm{A2} + \frac{1}{\mathcal{N}} \cdot \left(W_1 + W_2 + 2 \cdot \sqrt{W_1 W_2}\right) && (39)\\
\mathrm{A1} - \mathrm{A2} &> \frac{W_1 + W_2 + 2 \cdot \sqrt{W_1 W_2} - 2 \cdot W_1 - W_2}{\mathcal{N}} && (40)\\
\mathrm{A1} - \mathrm{A2} &> \frac{2 \cdot \sqrt{W_1 W_2} - W_1}{\mathcal{N}} && (41)
\end{aligned}$$

This condition is similar to the condition under which parallel processing is better than serial processing without the perceptual or motor stages, and in fact (a) the durations of the motor stages are irrelevant, and (b) the two conditions are identical when A1 = A2. Other things being equal, parallel processing tends to be favored when A1 is much larger than A2, because A1 contributes twice to the total time in the serial mode but only once in the parallel mode. In conclusion, then, the analysis of three-stage models also suggests that under common experimental conditions (i.e., $W_2$ is not too much larger than $W_1$), the total time for completing both tasks would be less with the serial mode than with the parallel mode.

## Appendix C

## Overlap-Dependent Models

Elsewhere in the article, we have for simplicity considered only parallel models in which the speed of parallel processing is independent of the temporal overlap of central processing for the two tasks. Within these models, for example, parallel-mode RTs do not depend on SOA or—in stochastic versions—on the finishing times of the perceptual processes. Thus, we will refer to these as "overlap-independent" parallel models. The purpose of this appendix is to consider the alternative class of parallel models in which parallel-mode RTs do depend on overlap. Naturally, these will be referred to as "overlap-dependent" models.

Overlap-dependent models can be motivated within the frameworks of both limited-capacity models and outcome conflict models. Within capacity models, RTs would depend on SOA if capacity could be flexibly reallocated within a trial. The most prominent examples of such models are limited-capacity parallel models in which capacity is shared only among the tasks that are ready for processing (e.g., Navon & Miller, 2002; Tombu & Jolicœur, 2003). In such models, tasks are processed more rapidly when they demand capacity at different times (e.g., if SOA is large), because processing resources can be concentrated on a single task when only that task is ready for processing (e.g., concentrated on each task separately when SOA is large). In contrast, the overlap-independent models considered in the main text are intuitively rather inefficient, because processing is no faster when only a single task is ready to be processed. In the carwash analogy, for example, the overlap-independent models correspond to the situation in which (a) the workers are divided into two teams of three, and (b) one team stands idle in the interval between the arrivals of the first and second cars. With overlap-dependent processing, however, both teams would work on the first car until the second car arrived, after which one team would be reallocated to work on that second car. Obviously, overlap-dependent limited-capacity parallel models would be more efficient than overlap-independent ones, and the purpose of this appendix is to consider whether serial processing would still be more efficient than overlap-dependent parallel processing.

Overlap-dependent models can also be motivated within the framework of outcome conflict

models. Within these models, dual-task interference is caused by outcome conflict or interfering crosstalk that arises when people must prepare to perform, or actually perform, two tasks at the same time. In essence, the overlap-independent models considered to this point suggest that the interfering effects of crosstalk are independent of overlap because they depend only on the readiness to perform both tasks, not on the actual performance of them both. If some of the interference arose only when both tasks were being performed simultaneously, however, then interference would depend on overlap. In particular, there would be less interference when overlap was small. Again, then, overlap-dependent outcome conflict models would suffer less interference than overlap-independent models, and they might therefore be more efficient than serial models.

---

Insert Figure C1 about here

---

Figure C1 illustrates the nature of parallel processing within overlap-dependent models. This figure is a generalization of Figure 5 of Tombu and Jolicœur (2003), and we have retained their labelling of Cases A–F. This figure generalizes theirs with respect to the rates at which central processing is carried out during parallel processing. They examined a fixed capacity model in which $r_2 = 1 - r_1$, but we consider the more general case in which these two rates are not strictly dependent in this fashion. This generalization could be motivated within limited-capacity models by assuming that there are some additional task-specific central capacities that do not need to be shared, in which case $r_2 > 1 - r_1$. Alternatively, it could be motivated within outcome conflict models by assuming that there is an arbitrary reduction of processing rate during task overlap, with the level of interference between tasks depending on the specific similarities between tasks. Corresponding to Figure C1, Table C1 indicates the relationships among stage processing durations that determine which case is the appropriate description of processing.

---

Insert Table C1 about here

---

*Comparison of Serial and Parallel Processing*

*Serial Processing.* When processing is serial, we assume that the central process is allocated to the first task to finish perceptual processing. Without loss of generality, we assume that the rate of processing in the serial mode is one unit per millisecond, so the duration of a stage is simply the amount of work required. Under that assumption, the RTs for the two tasks can be summarized as follows:

$$\mathrm{RT1}_s = \begin{cases} \mathrm{A1} + \mathrm{B1} + \mathrm{C1}, & \text{Case A, B, C, or F} \\ \mathrm{SOA} + \mathrm{A2} + \mathrm{B2} + \mathrm{B1} + \mathrm{C1}, & \text{Case D or E} \end{cases} \tag{42}$$

$$\mathrm{RT2}_s = \begin{cases} \mathrm{A2} + \mathrm{B2} + \mathrm{C2}, & \text{Case A, D, E, or F} \\ \mathrm{A1} + \mathrm{B1} + \mathrm{B2} + \mathrm{C2} - \mathrm{SOA}, & \text{Case B or C} \end{cases} \tag{43}$$

Note that $\mathrm{RT2}_s$ is prolonged by slack in cases B and C, whereas $\mathrm{RT1}_s$ is prolonged by slack in cases D and E.

*Cases A and F.* In these two cases, processing is always effectively serial. In case A, for example, $\mathrm{A1} + \mathrm{B1} < \mathrm{A2} + \mathrm{SOA}$, so central processing of Task 1 finishes before central processing of Task 2 is ready to start, even if the participant intends to use the parallel processing mode. Thus, for these cases the same RTs are predicted whether the participant intends to process in the serial or parallel mode—namely, the RTs predicted by the serial mode. The conclusion is that the parallel and serial processing modes are equivalent for these two cases.

*Case B.* When processing is parallel, the RTs for this case are

$$\mathrm{RT1}_p = \mathrm{SOA} + \mathrm{A2} + \frac{\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}}{r_1} + \mathrm{C1} \tag{44}$$

$$\mathrm{RT2}_p = \mathrm{A2} + \frac{\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}}{r_1} + \mathrm{B2} - r_2 \cdot \frac{\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}}{r_1} + \mathrm{C2} \tag{45}$$

Omitting the values of C1 and C2 that contribute to the total RTs equally under serial and parallel processing, it can be seen that serial processing is more efficient if

$$\mathrm{RT1}_s + \mathrm{RT2}_s < \mathrm{RT1}_p + \mathrm{RT2}_p \tag{46}$$

$$\begin{aligned} 2 \cdot \mathrm{A1} + 2 \cdot \mathrm{B1} + \mathrm{B2} - \mathrm{SOA} < {} & \mathrm{SOA} + \mathrm{A2} + \frac{\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}}{r_1} \\ & + \mathrm{A2} + \frac{\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}}{r_1} + \mathrm{B2} - r_2 \cdot \frac{\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}}{r_1} \end{aligned}$$

$$\begin{aligned}
2 \cdot \mathrm{A1} + 2 \cdot \mathrm{B1} - \mathrm{SOA} &< \mathrm{SOA} + 2 \cdot \mathrm{A2} + 2 \cdot \frac{\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}}{r_1} \\
&\quad - r_2 \cdot \frac{\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}}{r_1} \\
2 \cdot (\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}) &< (\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}) \cdot \left( \frac{2}{r_1} - \frac{r_2}{r_1} \right) \\
2 &< \left( \frac{2}{r_1} - \frac{r_2}{r_1} \right) \\
2 \cdot r_1 + r_2 &< 2 \qquad (47)
\end{aligned}$$

For example, in the limited-capacity model where $r_2 = 1 - r_1$, this inequality is always satisfied, so serial processing is always more efficient than parallel. In alternative models where $r_1 = r_2 = r$, serial processing is more efficient if $r < 2/3$.

*Case C.* When processing is parallel, the RTs for this case are

$$\mathrm{RT1}_p = \mathrm{A1} + \frac{\mathrm{B2}}{r_2} + \mathrm{B1} - \frac{\mathrm{B2}}{r_2} \cdot r_1 + \mathrm{C1} \qquad (48)$$

$$\mathrm{RT2}_p = \mathrm{A2} + \frac{\mathrm{B2}}{r_2} + \mathrm{C2} \qquad (49)$$

Again omitting the values of C1 and C2, serial processing is more efficient if

$$\begin{aligned}
2 \cdot \mathrm{A1} + 2 \cdot \mathrm{B1} + \mathrm{B2} - \mathrm{SOA} &< \mathrm{A1} + 2 \cdot \frac{\mathrm{B2}}{r_2} + \mathrm{B1} - \frac{\mathrm{B2}}{r_2} \cdot r_1 + \mathrm{A2} \\
\mathrm{A1} + \mathrm{B1} + \mathrm{B2} - \mathrm{SOA} - \mathrm{A2} &< 2 \cdot \frac{\mathrm{B2}}{r_2} - \frac{\mathrm{B2}}{r_2} \cdot r_1 \\
1 + \frac{\mathrm{A1} + \mathrm{B1} - \mathrm{SOA} - \mathrm{A2}}{\mathrm{B2}} &< \frac{2 - r_1}{r_2} \qquad (50)
\end{aligned}$$

Note that the fraction on the left side of this inequality is the ratio of the amount of slack to the duration of the postponed central stage. When the amount of slack is large relative to the duration of the postponed stage, parallel processing is more efficient than serial for a wider range of $r_1$ and $r_2$ values.

As an example for this case, in the limited-capacity model where $r_1 = 1 - r_2$, serial processing is better when $r_2 < \frac{\mathrm{B2}}{\mathrm{A1+B1-SOA-A1}}$. Note that the denominator on the right side of this inequality is the amount of slack, and that the inequality must be satisfied when the slack is less than B2—which must happen at large values of SOA—because $r_2 \leq 1$. As another example, in models where $r_1 = r_2 = r$, serial processing is more efficient if $r < 2/\left(2 + \frac{\mathrm{A1+B1-SOA-A2}}{\mathrm{B2}}\right)$. That is, serial processing will typically be more efficient than parallel processing when $r$ is smaller

than a cutoff of approximately 2/3 if SOA=0 and the tasks are approximately equivalent (i.e., A1 ≈ A2 and B1 ≈ B2). The cutoff increases with SOA.

*Case D.* This case is similar to Case C, except that Task 2 reaches the central processing stage before Task 1. When processing is parallel, the RTs for this case are

$$\mathrm{RT1}_p = \mathrm{A1} + \frac{\mathrm{B1}}{r_1} + \mathrm{C1} \tag{51}$$

$$\mathrm{RT2}_p = \mathrm{A2} + \frac{\mathrm{B1}}{r_1} + \mathrm{B2} - \frac{\mathrm{B1}}{r_1} \cdot r_2 + \mathrm{C2} \tag{52}$$

Again omitting the values of C1 and C2, serial processing is more efficient if

$$\begin{aligned} \mathrm{SOA} + 2 \cdot \mathrm{A2} + 2 \cdot \mathrm{B2} + \mathrm{B1} &< \mathrm{A1} + 2 \cdot \frac{\mathrm{B1}}{r_1} + \mathrm{A2} + \mathrm{B2} - \frac{\mathrm{B1}}{r_1} \cdot r_2 \\ \mathrm{SOA} + \mathrm{A2} + \mathrm{B2} + \mathrm{B1} - \mathrm{A1} &< \mathrm{B1} \cdot \frac{2 - r_2}{r_1} \\ 1 + \frac{\mathrm{SOA} + \mathrm{A2} + \mathrm{B2} - \mathrm{A1}}{\mathrm{B1}} &< \frac{2 - r_2}{r_1} \end{aligned} \tag{53}$$

This result is analogous to the result for Case C. Again, the fraction on the left side of this inequality is the ratio of the amount of slack to the duration of the postponed central stage, and the $r_1$ and $r_2$ terms are interchanged on the right side, relative to those in Inequality 50. Thus, serial processing is better than parallel processing under conditions analogous to those already discussed in connection with Case C.

*Case E.* This case is similar to Case B, except that Task 2 reaches the central processing stage before Task 1. When processing is parallel, the RTs for this case are

$$\mathrm{RT1}_p = \mathrm{A1} + \frac{\mathrm{B2} - (\mathrm{A1} - \mathrm{A2} - \mathrm{SOA})}{r_2} + \mathrm{B1} - r_1 \cdot \frac{\mathrm{B2} - (\mathrm{A1} - \mathrm{A2} - \mathrm{SOA})}{r_2} + \mathrm{C1} \tag{54}$$

$$\mathrm{RT2}_p = \mathrm{A1} - \mathrm{SOA} + \frac{\mathrm{B2} - (\mathrm{A1} - \mathrm{A2} - \mathrm{SOA})}{r_2} + \mathrm{C2} \tag{55}$$

Again omitting the values of C1 and C2, serial processing is more efficient if

$$\begin{aligned} \mathrm{SOA} + 2 \cdot \mathrm{A2} + 2 \cdot \mathrm{B2} + \mathrm{B1} &< 2 \cdot \mathrm{A1} + 2 \cdot \frac{\mathrm{A2} + \mathrm{B2} + \mathrm{SOA} - \mathrm{A1}}{r_2} + \mathrm{B1} - \mathrm{SOA} \\ &\quad - r_1 \cdot \frac{\mathrm{A2} + \mathrm{B2} + \mathrm{SOA} - \mathrm{A1}}{r_2} \\ 2 \cdot (\mathrm{SOA} + \mathrm{A2} + \mathrm{B2} - \mathrm{A1}) &< (\mathrm{SOA} + \mathrm{A2} + \mathrm{B2} - \mathrm{A1}) \cdot \frac{2 - r_1}{r_2} \\ 2 &< \frac{2 - r_1}{r_2} \end{aligned} \tag{56}$$

As in Case B, this inequality is always satisfied—and serial processing is always better—for limited-capacity models in which $r_2 = 1 - r_1$. Alternatively, if $r_2 = r_1 = r$, serial processing is better if $r < 2/3$.

*Conclusion*

The analysis of overlap-dependent parallel models is quite complex because of the need to distinguish among six different cases. Nonetheless, serial processing is at least as efficient as parallel processing under a wide variety of parameter values for every case, especially when SOA is long and when the parallel processing rates $r_1$ and $r_2$ are substantially smaller than the serial processing rate. The overall conclusion, then, is that serial processing is likely to be more efficient than overlap-dependent parallel processing, just as it is likely to be more efficient than overlap-independent parallel processing.

This overall conclusion can best be illustrated by considering Case B, which is clearly the most plausible case for most PRP situations. For this case, the serial mode is more efficient than the limited-capacity model (i.e., $r_2 = 1 - r_1$) with instantaneous redistribution of a fixed resource pool among active tasks (cf. Navon & Miller, 2002; Tombu & Jolicœur, 2003). Thus, even if people had available this relatively efficient parallel processing mode, they would still be advised to process in serial.

It should be emphasized that parallel models with instantaneous reallocation of capacity across tasks, as considered in this appendix, represent the "best case" for parallel processing in the sense that parallel models needing time to reallocate capacity would necessarily require longer total processing times to accomplish the same tasks. Having shown that serial processing is often more efficient than even these best-case parallel models, then, tends to generalize substantially the argument that serial processing is often more efficient than parallel processing. Consideration of this extreme case was arguably unnecessary, however, because typical RT1 data rule out optimal parallel models. Specifically, under these models RT1 should decrease as SOA increases, because a longer SOA lets Task 1 be processed longer with full capacity; this pattern is not, however, generally observed (Pashler, 1994a; but see Tombu & Jolicœur, 2002).

Although this appendix shows that serial processing is generally more efficient than parallel

processing even when the speed of parallel processing increases for nonoverlapping tasks, we are aware of one somewhat artificial example in which the two processing modes would be equally efficient.[10] For simplicity, we will describe this example in detail only for the case of a single processing stage for each task and for SOA=0. In the serial mode of this example, the time needed to perform each task is an exponential random variable with rate $2\lambda$. As a result, with SOA=0,

$$\begin{aligned} \mathrm{E}\left[\mathrm{RT1}_s\right] &= \frac{1}{2\lambda} \\ \mathrm{E}\left[\mathrm{RT2}_s\right] &= \frac{1}{2\lambda} + \frac{1}{2\lambda} = \frac{2}{2\lambda} \\ \mathrm{E}\left[\mathrm{TRT}_s\right] &= \mathrm{E}\left[\mathrm{RT1}_s\right] + \mathrm{E}\left[\mathrm{RT2}_s\right] = \frac{3}{2\lambda} \end{aligned}$$

In the parallel mode, each task is initially processed with rate $\lambda$. When one task finishes, however, the remaining task is thereafter processed with rate $2\lambda$ (cf. Townsend & Ashby, 1983, Chap. 4). As a result, with SOA=0,

$$\begin{aligned} \mathrm{E}\left[\min(\mathrm{RT1}_p, \mathrm{RT2}_p)\right] &= \frac{1}{2\lambda} \\ \mathrm{E}\left[\max(\mathrm{RT1}_p, \mathrm{RT2}_p)\right] &= \frac{1}{2\lambda} + \frac{1}{2\lambda} = \frac{2}{2\lambda} \\ \mathrm{E}\left[\mathrm{TRT}_p\right] &= \mathrm{E}\left[\mathrm{RT1}_p\right] + \mathrm{E}\left[\mathrm{RT2}_p\right] = \frac{3}{2\lambda} \end{aligned}$$

Therefore, the total time TRT is equal to $3/(2\lambda)$ at $SOA = 0$ for both processing modes, so the two modes would be equally efficient. Although the analysis is more complex, $\mathrm{TRT}_p = \mathrm{TRT}_s$ also holds for a three-stage version of this model in which prebottleneck and postbottleneck processes are unlimited-capacity processes with exponential finishing times.

The problematic result of $\mathrm{E}\left[\mathrm{TRT}_p\right] = \mathrm{E}\left[\mathrm{TRT}_s\right]$ depends critically on the assumption of exponentially-distributed processing times for the limited-capacity central stage. For example, suppose the performance of these stages is modeled as a gamma with shape parameter two (i.e., the central stage requires two successive exponential steps rather than just one). Furthermore, again assume that each task is processed with a rate of $\lambda$ while the other task is still in progress and with a rate of $2\lambda$ once the other task has finished. In that case, parallel processing is slower than serial processing (in fact, $\mathrm{E}\left[\mathrm{TRT}_p\right] = 1.083 \cdot \mathrm{E}\left[\mathrm{TRT}_s\right]$), in keeping with the main conclusion of this appendix. In practice, then, the isolated exception provided by the exponential example is not strongly problematic for the main conclusions of this appendix, because the exponential

distribution is rarely if ever a realistic model for the time needed to perform cognitive tasks. Moreover, as noted earlier, any time costs of reallocating attentional capacity from one task to the other would selectively slow parallel processing and thereby yield $TRT_p > TRT_s$, in keeping with our overall conclusions, even if the other assumptions of the exponential model were met.

## Appendix D

## Task Preparation Account

This appendix explores the issue of whether the effects of SOA distribution on RT1 and especially RT2 can be explained by the bottleneck model if it is augmented by the concept of task preparation (e.g., Pashler, 1994a; cf. Navon & Miller, 2002). Within this augmented model, the two tasks are always processed in serial whether short or long SOAs are frequent, in accordance with the simple bottleneck model. It is assumed, however, that the SOA distribution influences the relative levels of preparation for Tasks 1 and 2. Specifically, there is assumed to be a tradeoff in the levels of preparation for the two tasks, from one theoretical extreme of being completely prepared for Task 1 and not at all prepared for Task 2, to the reverse extreme of being equally prepared for both tasks. Naturally, the amount of time needed for the bottleneck stage to process a given task is assumed to decrease as the level of preparation for that task increases. Finally, it is assumed that relative preparation is influenced by the distribution of SOAs: When SOA is usually long, participants would tend to prepare mostly for Task 1 and relatively little for Task 2; but when SOA is usually short, participants would tend to prepare more equally for the two tasks.

The question is whether this bottleneck model augmented with the idea of differential preparation could account for the observed patterns of RTs. In particular we focus on the issue of whether it could account for the observed crossovers of the functions relating RT2 to SOA for the SF and LF conditions of Experiments 1–3 (cf. Figures 3, 5, and 7).

First, note that according to the three-stage bottleneck model (e.g., Pashler & Johnston, 1989), at small values of SOA+$A1$, RT2 is given by

$$\mathrm{RT2} = A1 + B1 + B2 + C2 - SOA \tag{57}$$

and at long values of SOA it is

$$\mathrm{RT2} = A2 + B2 + C2. \tag{58}$$

$B1$ and $B2$ represent the durations of the bottleneck processes for Tasks 1 and 2, respectively. Furthermore, $A1$ and $A2$ denote the corresponding pre-bottleneck processes, and $C1$ and $C2$ denote the corresponding post-bottleneck processes.

Second, let $RT2_{SF}$ and $RT2_{LF}$ denote the RT2's in conditions SF and LF, respectively. Note that the observed crossover of the functions relating RT2 to SOA implies the following inequalities: At short values of SOA

$$RT2_{LF} > RT2_{SF}, \tag{59}$$

whereas at long values of SOA

$$RT2_{LF} < RT2_{SF}. \tag{60}$$

Third, let $B1_{SF}$ and $B2_{SF}$ be the durations of the bottleneck processes for Tasks 1 and 2, respectively, in condition SF. Analogously, let $B1_{LF}$ and $B2_{LF}$ be the durations of these processes in condition LF. With these definitions, inequality 60 can be rewritten as

$$A2 + B2_{LF} + C2 < A2 + B2_{SF} + C2 \tag{61}$$

$$B2_{LF} < B2_{SF} \tag{62}$$

$$B2_{SF} - B2_{LF} > 0, \tag{63}$$

and inequality 59 can be rewritten as

$$A1 + B1_{LF} + B2_{LF} + C2 - SOA > A1 + B1_{SF} + B2_{SF} + C2 - SOA \tag{64}$$

$$B1_{LF} + B2_{LF} > B1_{SF} + B2_{SF}. \tag{65}$$

These two inequalities can be combined into the overall inequality

$$B1_{LF} - B1_{SF} > B2_{SF} - B2_{LF} > 0, \tag{66}$$

and the parameter values satisfying this inequality are exactly those for which the bottleneck model would predict a crossover of the RT2 functions in conditions SF versus LF.

Figure D1 illustrates the predictions of this augmented bottleneck model. It shows that the model can produce a crossover when the values of $B1_{SF}$, $B1_{LF}$ $B2_{SF}$, and $B2_{LF}$ satisfy inequality 66. At short values of SOA, RT2 is larger in condition LF than in condition SF. In both conditions, RT2 decreases as SOA increases, with a slope of -1, just as it did in the simple bottleneck model. (Indeed, inspection of Equation 57 indicates that the bottleneck model must always produce a slope of -1 within any condition, regardless of the effects of relative preparation

on $B1$ and $B2$ within that condition.) Nonetheless, a crossover emerges because the RT2 function in condition SF levels off at a value of SOA for which the RT2 function is still decreasing in condition LF. Thus, at a qualitative level this model can produce the crossover of RT2 functions in the SF and LF conditions observed in our experiments. In addition, the figure also displays the predictions for RT1 in both SOA conditions, and these are qualitatively in accord with the data from Experiments 1–3.

---

Insert Figure D1 about here

---

Although the augmented bottleneck model can produce a crossover interaction of SOA and SOA distribution in the direction observed in the present experiments, we conclude that this model cannot actually provide a plausible account for the results. There are two main reasons for this conclusion. The first one is that there are quantitative discrepancies between the predictions of the augmented bottleneck model and the observed data concerning the initial slopes of the RT2 functions. Note that the model requires the RT2 functions to have equal -1 slopes over the range of small SOAs in both conditions, SF and LF. The model produces a crossover only because of differences in the asymptotes of the SOA functions. The observed data, however, suggest that the slopes are different in the two conditions even for the smallest SOAs (i.e., with SOA in the range of 16–133 ms; cf. Figures 3 and 5). In addition, the model predicts that the crossover point should occur after one function has reached its asymptote, but the data indicate that the crossover occurs well above the asymptotic level for both functions. Furthermore, simulations show that the model still makes these two predictions (i.e., -1 slope and crossover at asymptote) even if there is random variability in the finishing times of stages. With moderate variability (i.e., coefficient of variation = 0.1; cf. Luce, 1986), for example, the -1 slopes of both functions shown in Figure D1 are maintained out to very nearly the asymptotic SOA of the SF function, and the crossover occurs within less than 5 ms of the asymptote. Even with unrealistically large variability (i.e., coefficient of variation = 0.3), the -1 slope is maintained out to an SOA of over 100 ms and the crossover occurs within approximately 15 ms of the asymptote. Thus, it does not appear that the bottleneck model augmented by differential task preparation can provide an accurate quantitative

account for the observed effects of SOA distribution on the slope of RT2, even with stochastic process durations.

The second reason for concluding that the augmented bottleneck model cannot account for the results is that the parameter values needed for it to produce the observed crossover are psychologically quite implausible. As is summarized in inequality 66, the observed crossover can only be obtained when both $B1_{LF} - B1_{SF}$ and $B2_{SF} - B2_{LF}$ are greater than zero. Consideration of the most likely effects of task preparation, however, suggests that these differences are exactly the opposite of what would be expected. Consider first preparation for Task 2: It should be greater in the SF condition than in the LF condition, because participants should prepare more equally for the two tasks when SOAs are usually short. Assuming that Task 2 bottleneck processing time decreases with increasing preparation for Task 2, it follows that $B2_{SF}$ should be less than $B2_{LF}$. Thus, the most plausible account in terms of task preparation suggests that $B2_{SF} - B2_{LF}$ should be negative—not positive as is required for the model to predict the crossover interaction. Similarly, preparation for Task 1 should be greater in the LF condition than in the SF condition, implying that $B1_{LF} - B1_{SF}$ should also be negative—also in the wrong direction to produce the observed interaction.

## Author Note

We thank Eric Ruthruff for information about response grouping, David Meyer and Richard Schweickert for helpful comments on a previous version of this article, and Isabelle Schurr for her assistance in collecting the data. Correspondence concerning this article should be addressed to Jeff Miller, Department of Psychology, University of Otago, Dunedin, New Zealand or Rolf Ulrich, Abteilung für Allgemeine und Biologische Psychologie, Psychologisches Institut, Universität Tübingen, Friedrichstr. 21, 72072 Tübingen, Germany. Electronic mail may be sent to miller@psy.otago.ac.nz or ulrich@uni-tuebingen.de. This research was supported by a grant (Ul 116/6-2) from the Deutsche Forschungsgemeinschaft to Rolf Ulrich and by a grant to Jeff Miller from the Marsden Fund administered by the Royal Society of New Zealand.

**Footnotes**

[1] To simplify the discussion, we concentrate on cognitive rather than peripheral factors, ignoring the fact that structural or anatomical limitations sometimes prevent parallel processing of two tasks. For example, people cannot foveate two different locations in visual space in parallel, nor can they simultaneously reach with the right hand toward two different response manipulanda.

[2] The dichotomy of serial versus parallel processing is a convenient simplification, but the true range of theoretical possibilities is much more complex than that. For example, processing capacity might be divided between tasks with the proportions of 90% and 10% rather than with the 50%/50% split associated with maximally parallel processing or with the 100%/0% split associated with maximally serial processing. As this example illustrates, participants can in principle adopt any of a potentially limitless number of intermediate modes by adjusting the relative priorities of the two tasks (e.g., Navon & Miller, 2002; Tombu & Jolicœur, 2003). In this respect, the degree of serial versus parallel processing varies quantitatively rather than dichotomously, so it is appropriate to speak of processing as being "more serial" or "more parallel". A second complication is that participants might use a serial mode in some proportion of trials and a parallel mode in the rest of the trials, producing a probability mixture of the two modes across a full set of trials. In this case, processing could be said to be "more often serial" or "more often parallel". To acknowledge the possibility that the mode of processing can shift in a potentially graded fashion between the serial and parallel extremes, we will often refer to processing with terms suggesting a quantitative dimension from the serial extreme to the parallel one.

[3] In accordance with the canonical bottleneck model (see, e.g., Ruthruff, Pashler, & Hazeltine, 2003; Tombu & Jolicœur, 2003), the serial processing model includes no time cost for switching from Task 1 to Task 2. Although task-switching produces large time costs in many paradigms where two different tasks use the same stimulus sets (e.g., number stimuli for which the first task is to subtract three and the second task is to add six), such costs are much smaller or reversed in tasks using disjoint sets of stimuli (e.g., number stimuli to which three is added and

word stimuli to which an antonym is to be given; Jersild, 1927; Spector & Biederman, 1976). Because most PRP tasks use disjoint stimulus sets, the assumption of negligible task-switching cost seems to be a reasonable approximation for the present purposes.

[4] In some models the times needed for parallel processing might be less at the long SOA than at the short one, contrary to the situation depicted in this figure. In capacity models, for example, a plausible alternative assumption would be that processing resources are reallocated to Task 2 after the processing of Task 1 has completed. In that case, the overall task time would be $\mathrm{TRT} = X_p + X_s$. Although this lessens the disadvantage associated with the parallel mode, it does not eliminate it. It is even more favorable for parallel models to assume that processing capacity is only divided when two tasks compete for it at the same time, with full capacity devoted to a single task when that task is the only one that needs to be processed. With that assumption, at long SOAs both tasks are processed at full capacity, and the overall task time is $\mathrm{TRT} = 2 \cdot X_s$ just as in the serial model. As shown in Appendix C, however, serial processing still tends to be more efficient than processing even with this very favorable assumption about the allocation of capacity within a parallel model.

[5] Instructions vary somewhat across PRP studies, with most emphasizing first-task performance but some placing equal emphasis on both tasks (e.g., Pashler, 1994b; Ruthruff, Pashler, & Klaassen, 2001). We used equal-emphasis instructions because these seemed most likely to encourage participants to minimize total RT.

[6] To maximize power, trials with IRIs less than 100 ms were included rather than excluded from these analyses. A median IRI was computed separately for each participant in each condition, and the trials from that condition were then partitioned into those with IRIs shorter versus longer than the median.

[7] Because IRI is not independent of RT1 and RT2 but instead is derived from them (i.e., IRI = RT2 - RT1 + SOA), experimental effects on IRI can also be viewed as resulting directly from effects on RT1 and RT2. In the present instance, then, the reduction in IRI for the SF condition can also be viewed as consistent with the current account of the hypothesized changes in RT1 and RT2. Specifically, because processing tends to be more parallel in the SF condition—which increases RT1 and decreases RT2—it is natural for IRI to decrease in this condition as well.

[8] This account was suggested to us by Werner Sommer.

[9] Ruthruff, Pashler, and Klaassen (2001) extended the approach of Pashler (1994b) by presenting stimuli simultaneously (i.e., SOA = 0) in all trials and by requiring participants to group their responses. The use of all zero SOAs to encourage parallel processing is clearly quite a good idea in terms of the optimization framework. Moreover, requiring grouped responses is plausible intuitively, because grouping the responses seems to give maximal opportunity for processing both tasks at the same time. Unfortunately, their experimental comparison only allowed them to reject *unlimited-capacity* parallel processing. Specifically, they showed that RTs for a harder task depended on the difficulty of an easier task done at the same time, which they noted is compatible with limited-capacity parallel models as well as with serial ones.

[10]This example was pointed out to us by Wolfgang Schwarz.

Table 1

*Optimal Scheduling Mode as a Function of Processing Times and Distribution of Stimulus Onset Asynchronies (SOAs)*

| Relation of | Distribution of SOAs | |
|---|---|---|
| Processing Times | Short Frequent | Long Frequent |
| Serial Processing Fast ($X_p > 1.5 \cdot X_s$) | Serial | Serial |
| Parallel Processing Fast ($X_p < 1.5 \cdot X_s$) | Parallel | Serial |

Table 2

*Number of Trials per Block at Each Stimulus Onset Asynchrony (SOA) in Experiment 1*

| | SOA in ms | | | |
|---|---|---|---|---|
| Condition | 16 | 133 | 500 | 1000 |
| Short SOAs frequent (SF) | 192 | 144 | 96 | 48 |
| Long SOAs frequent (LF) | 48 | 96 | 144 | 192 |

Table B1

*Illustration of Processing Times Needed Under Capacity Model for Two Examples Differing in Amount of Work Required for Task 2.*

| | Processing Mode | | |
|---|---|---|---|
| | | Parallel | |
| | Serial | Equal Division | Optimal Division |
| *Example 1:* $W_1 = 9$, $W_2 = 16$ | | | |
| RT1 | 9 | 18 | 21 |
| RT2 | 25 | 32 | 28 |
| TRT | 34 | 50 | 49 |
| *Example 2:* $W_1 = 9$, $W_2 = 1$ | | | |
| RT1 | 9 | 18 | 12 |
| RT2 | 10 | 2 | 4 |
| TRT | 19 | 20 | 16 |

*Note.* Predicted reaction times (RT) for Tasks 1 and 2 and total RT (TRT) as a function of processing mode for two examples of task pairs requiring the indicated amounts of work $W$ for each task. For both examples, total capacity $\mathcal{N}$ was assumed to be 1. In example 1, the optimal division is to allocate 42.9% of capacity to Task 1; in example 2, this optimal proportion is 75%.

Table C1

*Conditions Determining Which of the Cases in Figure C1 Describes Processing.*

| | Conditions | | |
|---|---|---|---|
| Case | Which starts first? | Any overlap? | Which finishes first? |
| A | 1 | No | 1 |
| | $A1 \leq SOA + A2$ | $A1 + B1 \leq SOA + A2$ | Other conditions sufficient |
| B | 1 | Yes | 1 |
| | $A1 \leq SOA + A2$ | $A1 + B1 > SOA + A2$ | $\frac{A1+B1-SOA-A2}{r_1} \leq \frac{B2}{r_2}$ |
| C | 1 | Yes | 2 |
| | $A1 \leq SOA + A2$ | $A1 + B1 > SOA + A2$ | $\frac{A1+B1-SOA-A2}{r_1} > \frac{B2}{r_2}$ |
| D | 2 | Yes | 1 |
| | $A1 > SOA + A2$ | $A1 < SOA + A2 + B2$ | $\frac{A2+B2+SOA-A1}{r_2} > \frac{B1}{r_1}$ |
| E | 2 | Yes | 2 |
| | $A1 > SOA + A2$ | $A1 < SOA + A2 + B2$ | $\frac{A2+B2+SOA-A1}{r_2} \leq \frac{B1}{r_1}$ |
| F | 2 | No | 2 |
| | $A1 > SOA + A2$ | $A1 \geq SOA + A2 + B2$ | Other conditions sufficient |

*Note.* The six cases can be distinguished according to three conditions: (a) Which task, 1 or 2, starts first at the level of central processing, (b) Whether the tasks overlap at the level of central processing, and (c) Which task finishes first at the level of central processing. The status of each condition is shown on the same line as the case, and the stage durations necessary to produce that status are shown below each status indication. Further details are provided in the text.

## Figure Captions

*Figure 1.* Illustration of the optimization framework for the analysis of performance in psychological refractory period tasks. The stimulus onset asynchrony (SOA) separating the task onsets is either very short (i.e., SOA = 0) or long enough for the first task to be finished before the second task starts. $X_s$ and $X_p$ are the times needed for processing each task in the serial and parallel mode, respectively. TRT is the total reaction time summed across both tasks.

*Figure 2.* Illustrative predicted reaction times for the first and second tasks (RT1 and RT2) as a function of stimulus onset asynchrony (SOA) and of the probability distribution of SOAs. SF and LF indicate blocks of trials in which short versus long SOAs are frequent, respectively. The upper and lower panels illustrate possible smaller and larger effects of the SOA distribution, respectively, depending on the proportion of trials in which parallel processing is used in the SF condition. Predicted values of RT1 were computed using the equation $\text{RT1} = (1 - c) \cdot X_s + c \cdot X_p$. $X_s$ and $X_p$ are the times needed for processing in the serial and parallel modes, respectively. $c$ is the probability of processing in the parallel mode, which should be modulated by the distribution of SOAs, and $1 - c$ is the corresponding probability of processing in the serial mode. Predicted values of RT2 depend on SOA. For SOA = 0, the prediction is $\text{RT2} = (1 - c) \cdot 2 \cdot X_s + c \cdot X_p$. For the long SOA, however, the predictde value is $\text{RT2} = (1 - c) \cdot X_s + c \cdot X_p$. Predictions were computed using values of $X_p$ and $X_s$ equal to 280 ms and 200 ms, respectively. For the LF condition, $c = .1$ was used; for the SF condition, $c = .3$ was used for the upper panel and $c = .5$ was used for the lower panel.

*Figure 3.* Mean reaction times for Task 1 and 2, RT1 and RT2 (top panel); percentages of trials in which both responses were correct (second panel); percentages of trials with an interresponse interval (IRI) less than 100 ms (third panel); and mean trial-to-trial correlation between RT1 and RT2 (bottom panel), as a function of stimulus onset asynchrony (SOA) and SOA distribution in Experiment 1. SF and LF indicate SOA distributions with short SOAs frequent and long SOAs frequent, respectively.

*Figure 4.* Cumulative probability distributions of interresponse intervals as a function of stimulus

onset asynchrony (SOA) and SOA distribution in Experiment 1. SF and LF indicate SOA distributions with short SOAs frequent and long SOAs frequent, respectively.

*Figure 5.* Mean reaction times for Task 1 and 2, RT1 and RT2 (top panel); percentages of trials in which both responses were correct (second panel); percentages of trials with an interresponse interval (IRI) less than 100 ms (third panel); and mean trial-to-trial correlation between RT1 and RT2 (bottom panel), as a function of stimulus onset asynchrony (SOA) and SOA distribution in Experiment 2. SF and LF indicate SOA distributions with short SOAs frequent and long SOAs frequent, respectively.

*Figure 6.* Cumulative probability distributions of interresponse intervals as a function of stimulus onset asynchrony (SOA) and SOA distribution in Experiment 2. SF and LF indicate SOA distributions with short SOAs frequent and long SOAs frequent, respectively.

*Figure 7.* Mean reaction times for Task 1 and 2, RT1 and RT2 (top panel); percentages of trials in which both responses were correct (second panel); percentages of trials with an interresponse interval (IRI) less than 100 ms (third panel); and mean trial-to-trial correlation between RT1 and RT2 (bottom panel), as a function of stimulus onset asynchrony (SOA) and SOA distribution in Experiment 3. SF and LF indicate SOA distributions with short SOAs frequent and long SOAs frequent, respectively.

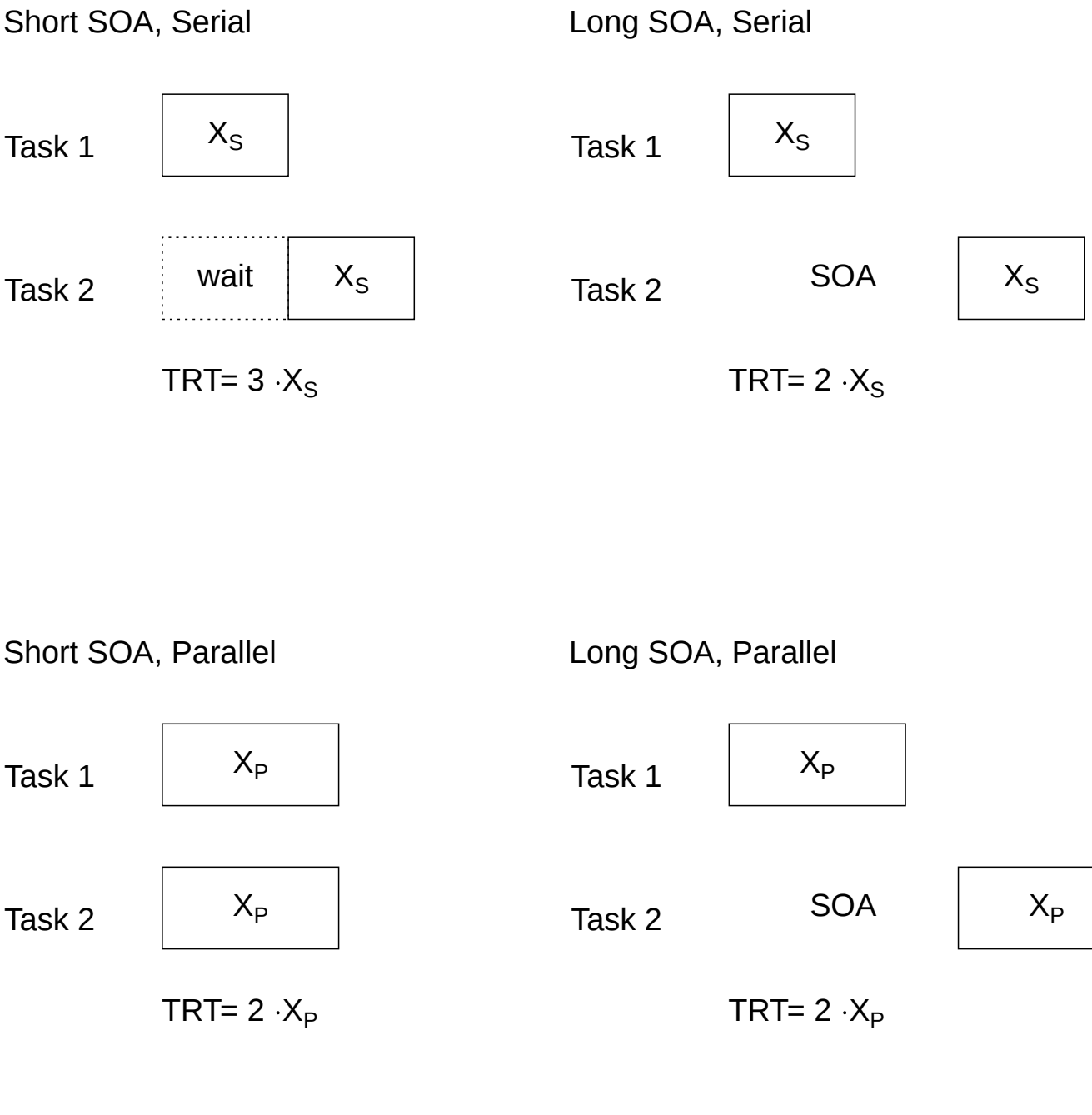
*Figure 8.* Cumulative probability distributions of interresponse intervals as a function of stimulus onset asynchrony (SOA) and SOA distribution in Experiment 3. SF and LF indicate SOA distributions with short SOAs frequent and long SOAs frequent, respectively.

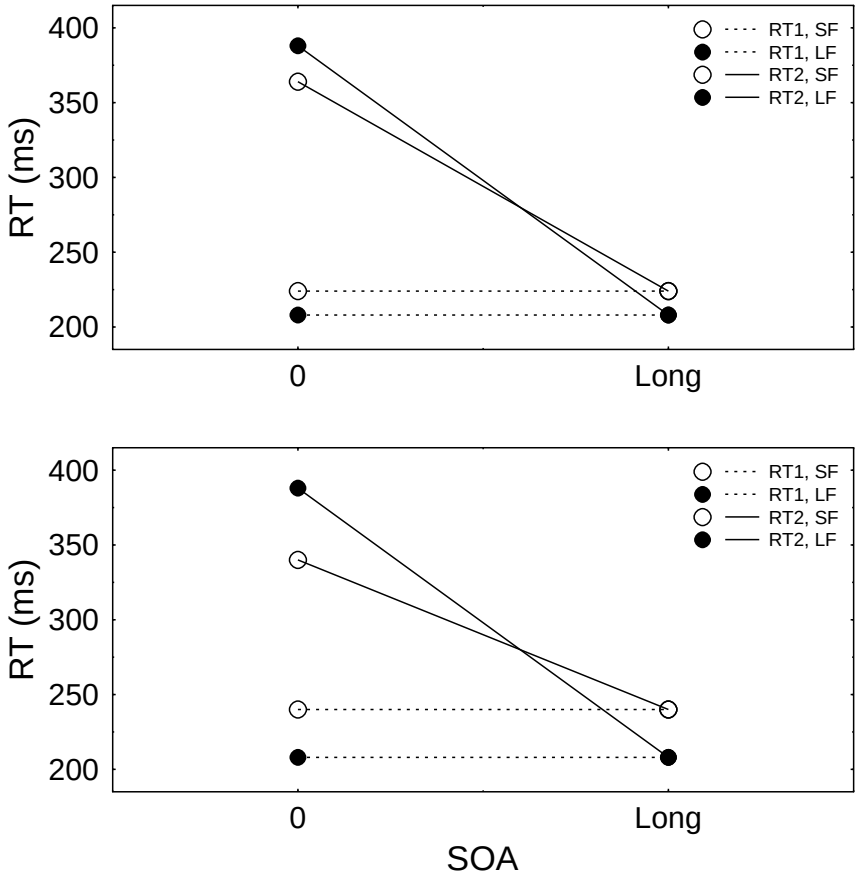
*Figure A1.* Illustration of the metatheoretical model for optimal scheduling of three-stage processes.

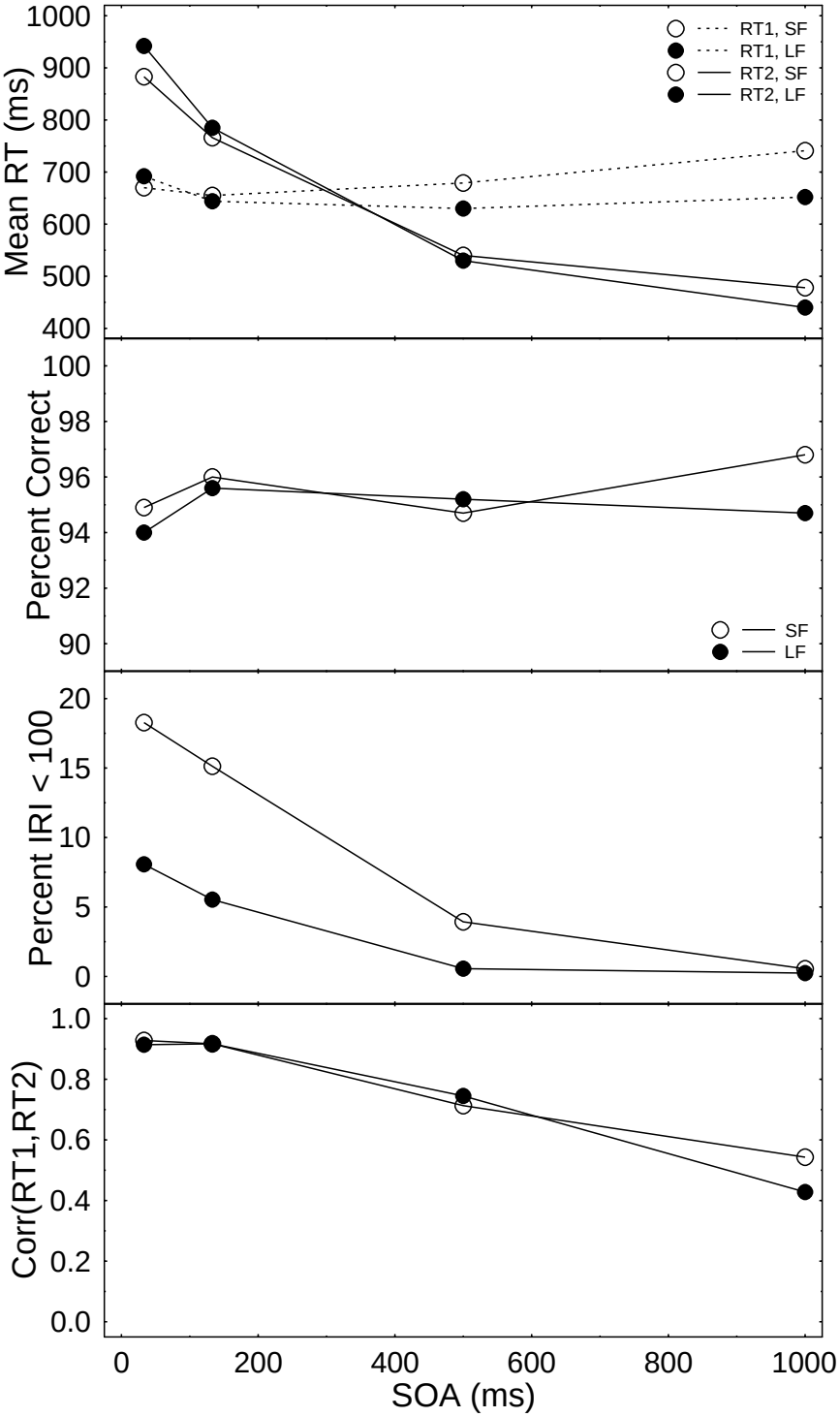
*Figure C1.* Depiction of dual-task processing for Tasks 1 and 2 within overlap-dependent parallel models. A1, B1, and C1 represent the perceptual, central, and motor processes for Task 1, respectively, and A2, B2, and C2 represent the corresponding processes for Task 2. The height of each process at each moment corresponds to its instantaneous processing rate. Note that A1, A2,

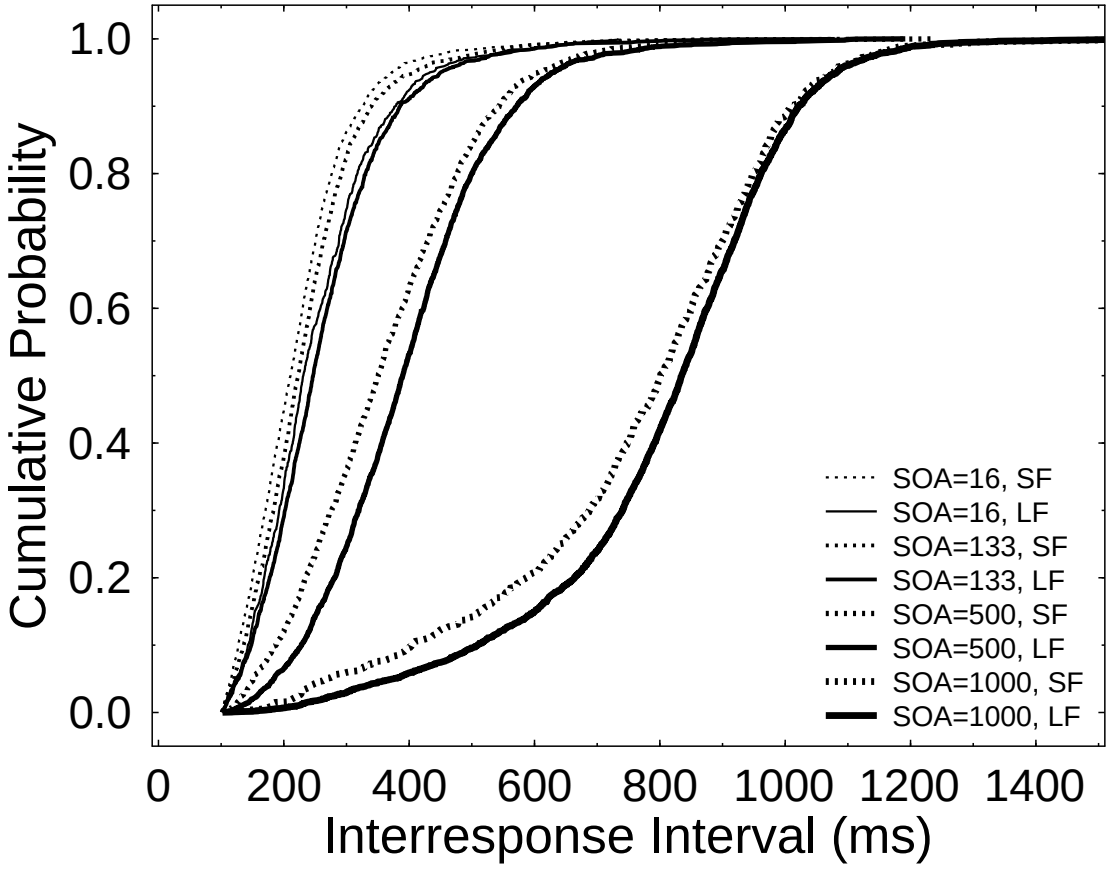
C1, and C2 operate at the same rates—depicted for simplicity as being equal—regardless of task overlap. In contrast, B1 and B2 operate at lower rates, $r_1$ and $r_2$, when they are both in operation simultaneously, so both take longer when they overlap. The total area of each process represents the amount of work needed for its completion; thus, when the work rates of B1 and B2 decrease, the total time needed for each to complete increases as needed to produce the same total work (i.e., area). The time needed for a specific process to complete (e.g., A1) can vary due to both experimental manipulations affecting the amount of work to be done and random trial-to-trial variations in this amount. This figure is modelled after Figure 5 of Tombu and Jolicœur (2003).

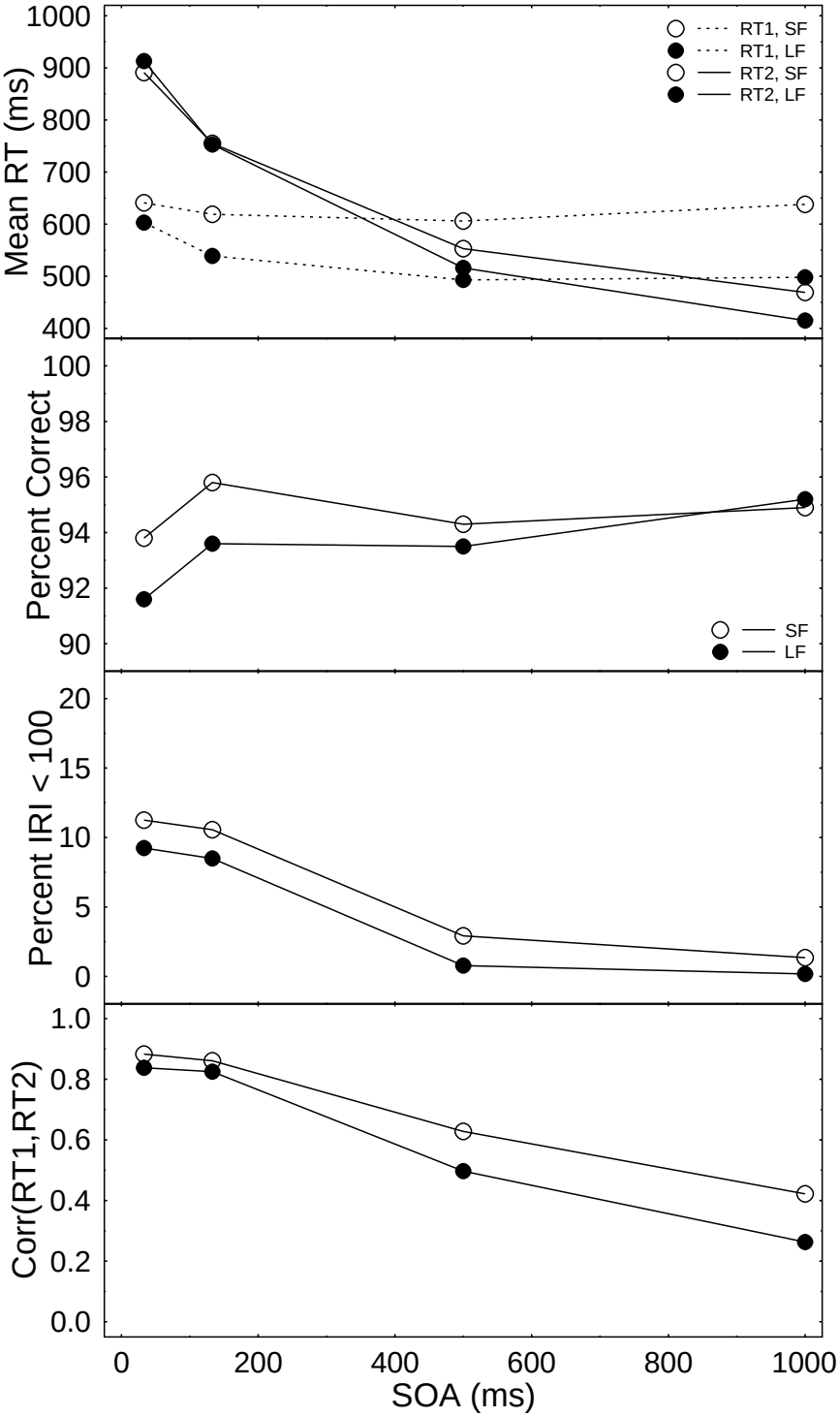
*Figure D1.* Illustration of predictions of the bottleneck model augmented with the assumption that task preparation depends on the distribution of SOAs. The parameter values were chosen to produce a crossover in the observed direction [i.e., RT2 decreases more across SOAs in the condition with long SOAs frequent (LF) than in the condition with short SOAs frequent (SF)]. The parameter values—all in milliseconds—were: $A1 = A2 = 150$, $C1 = C2 = 100$ $B1_{SF} = 180$, $B1_{LF} = 280$ $B2_{SF} = 250$, and $B2_{LF} = 200$. Notice that the values of $B1_{SF}$, $B1_{LF}$ $B2_{SF}$, and $B2_{LF}$ satisfy inequality 66.

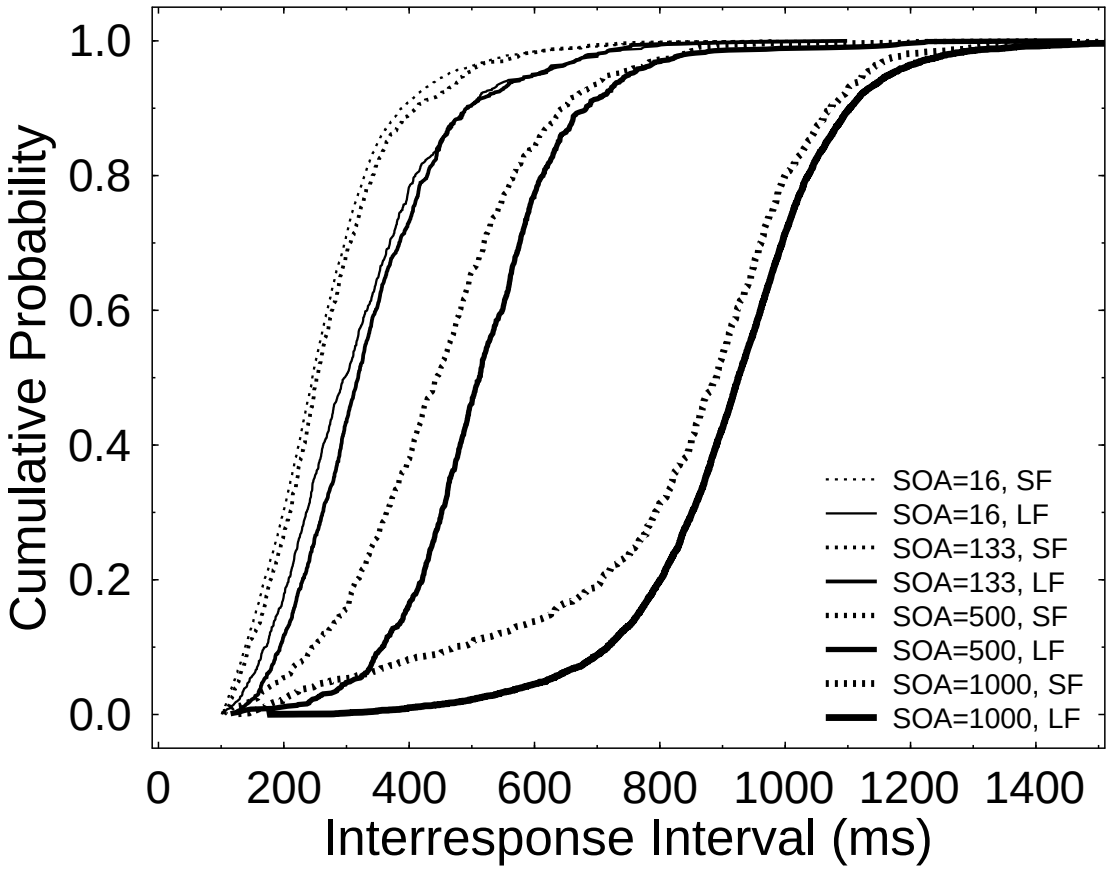

Short SOA, Serial
Task 1
$X_S$
Task 2
wait
$X_S$
TRT= 3 ·$X_S$
Long SOA, Serial
Task 1
$X_S$
Task 2
SOA
$X_S$
TRT= 2 ·$X_S$
Short SOA, Parallel
Task 1
$X_P$
Task 2
$X_P$
TRT= 2 ·$X_P$
Long SOA, Parallel
Task 1
$X_P$
Task 2
SOA
$X_P$
TRT= 2 ·$X_P$

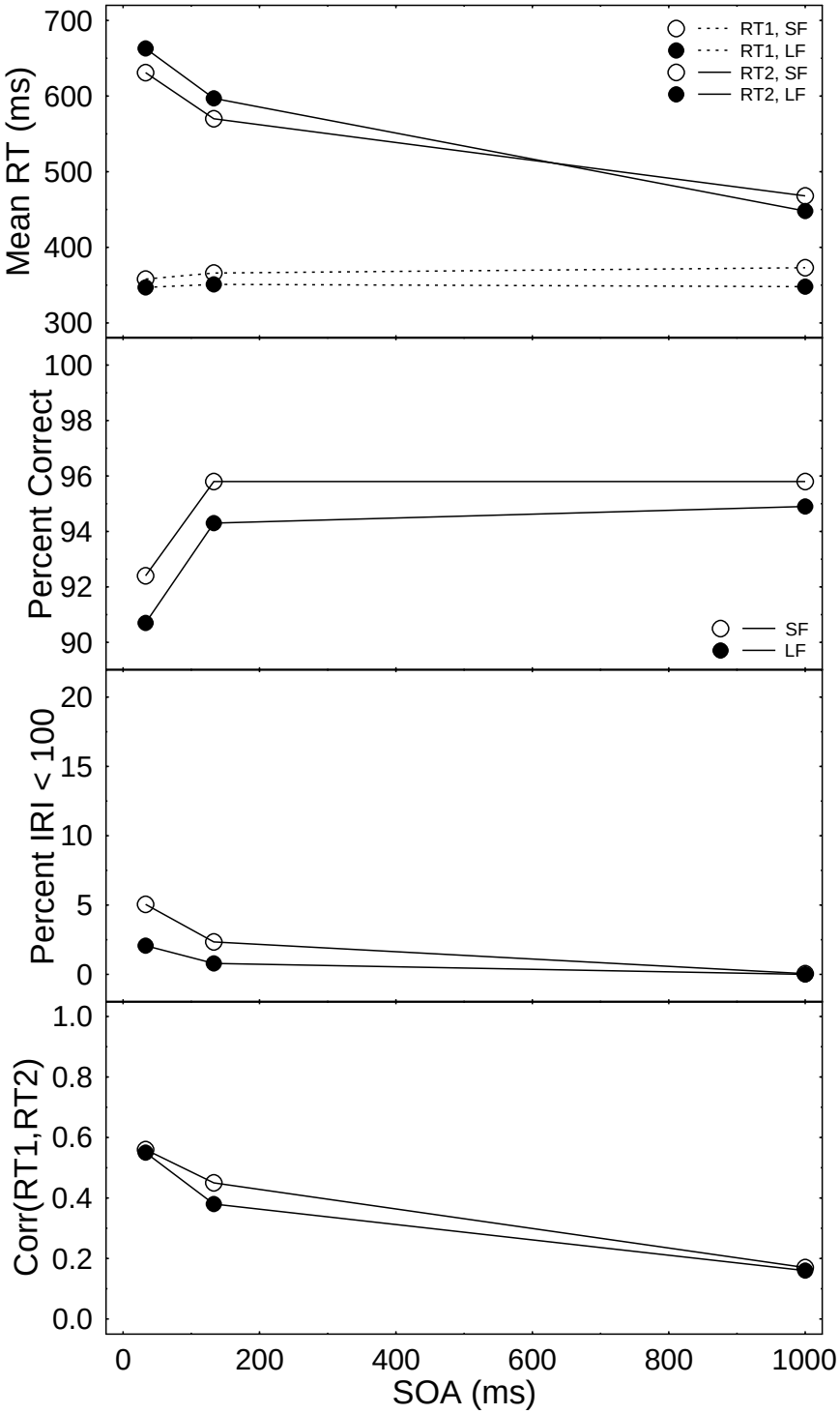

400
350
300
250
200
RT (ms)
RT1, SF
RT1, LF
RT2, SF
RT2, LF
0
Long
400
350
300
250
200
RT (ms)
RT1, SF
RT1, LF
RT2, SF
RT2, LF
0
Long
SOA

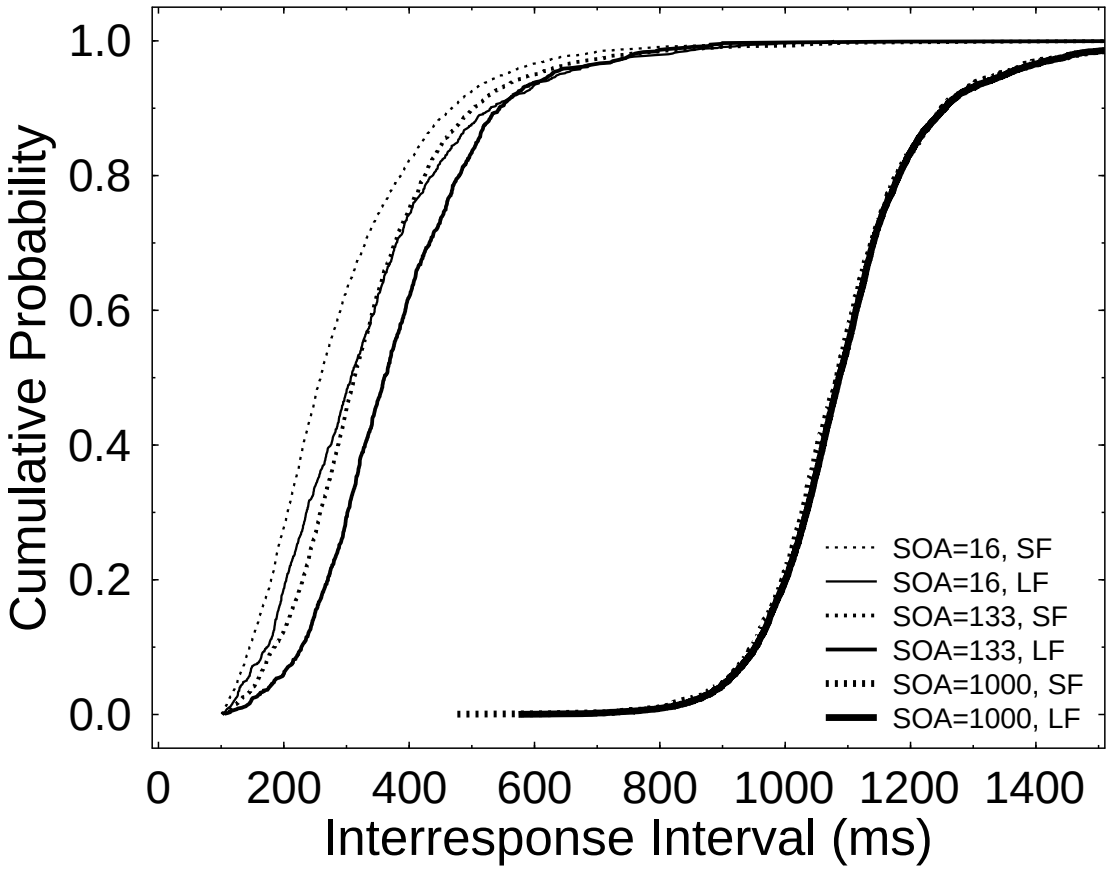

Mean RT (ms)
1000
900
800
700
600
500
400
RT1, SF
RT1, LF
RT2, SF
RT2, LF
Percent Correct
100
98
96
94
92
90
SF
LF
Percent IRI < 100
20
15
10
5
0
Corr(RT1,RT2)
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
SOA (ms)

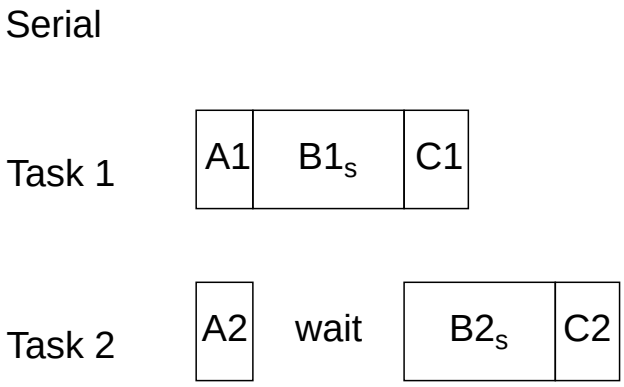

Cumulative Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200
1400
Interresponse Interval (ms)
SOA=16, SF
SOA=16, LF
SOA=133, SF
SOA=133, LF
SOA=500, SF
SOA=500, LF
SOA=1000, SF
SOA=1000, LF

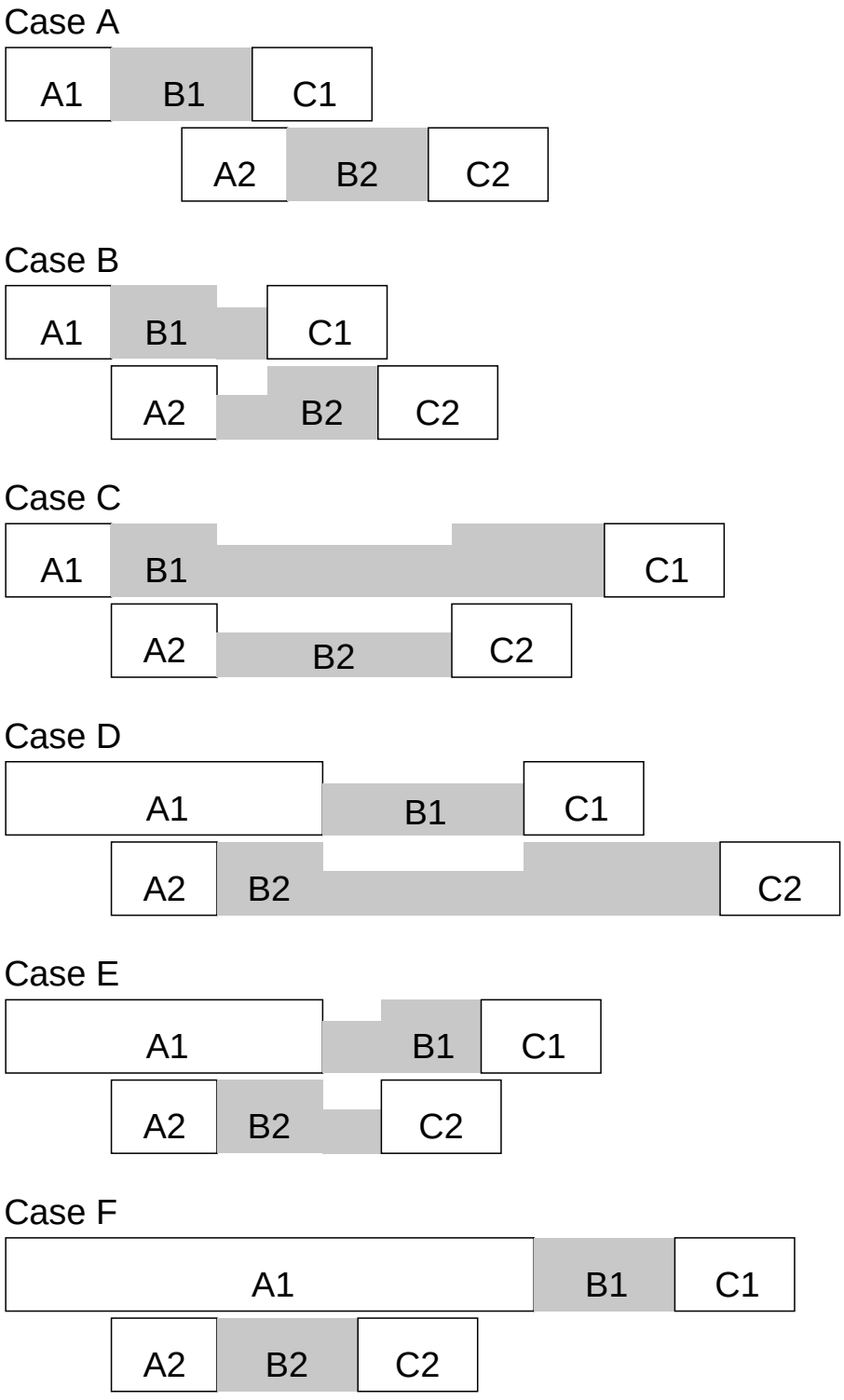

Mean RT (ms)
1000
900
800
700
600
500
400
RT1, SF
RT1, LF
RT2, SF
RT2, LF
Percent Correct
100
98
96
94
92
90
SF
LF
Percent IRI < 100
20
15
10
5
0
Corr(RT1,RT2)
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
SOA (ms)

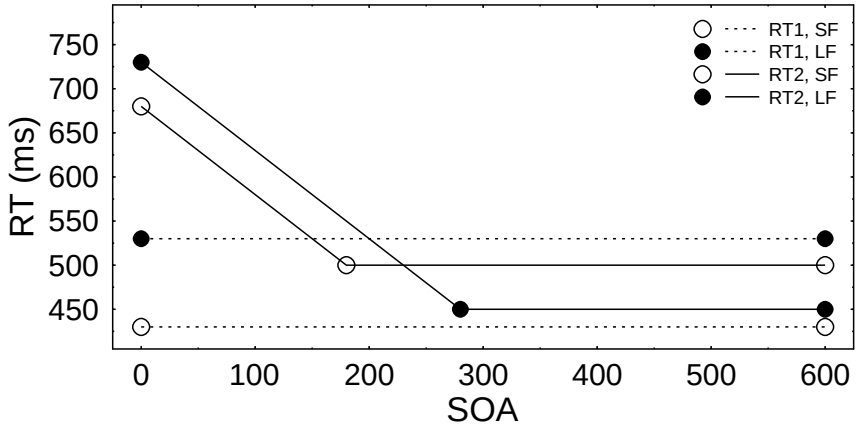

Cumulative Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200
1400
Interresponse Interval (ms)
SOA=16, SF
SOA=16, LF
SOA=133, SF
SOA=133, LF
SOA=500, SF
SOA=500, LF
SOA=1000, SF
SOA=1000, LF


Mean RT (ms)
700
600
500
400
300
RT1, SF
RT1, LF
RT2, SF
RT2, LF
Percent Correct
100
98
96
94
92
90
SF
LF
Percent IRI < 100
20
15
10
5
0
Corr(RT1,RT2)
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
SOA (ms)


Cumulative Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200
1400
Interresponse Interval (ms)
SOA=16, SF
SOA=16, LF
SOA=133, SF
SOA=133, LF
SOA=1000, SF
SOA=1000, LF


Serial
Task 1
A1
B1s
C1
Task 2
A2
wait
B2s
C2
Parallel
Task 1
A1
B1p
C1
Task 2
A2
B2p
C2


Case A
A1
B1
C1
A2
B2
C2
Case B
A1
B1
C1
A2
B2
C2
Case C
A1
B1
C1
A2
B2
C2
Case D
A1
B1
C1
A2
B2
C2
Case E
A1
B1
C1
A2
B2
C2
Case F
A1
B1
C1
A2
B2
C2


RT1, SF
RT1, LF
RT2, SF
RT2, LF
RT (ms)
750
700
650
600
550
500
450
0
100
200
300
400
500
600
SOA