Cognitive Psychology 58 (2009) 273–310

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

On the optimality of serial and parallel processing in the psychological refractory period paradigm: Effects of the distribution of stimulus onset asynchronies

Jeff Miller [a,*], Rolf Ulrich [b], Bettina Rolke [b]

[a] *Department of Psychology, University of Otago, Dunedin, New Zealand*
[b] *Abteilung für Allgemeine und Biologische Psychologie, Psychologisches Institut, Universität Tübingen, Friedrichstr. 21, 72072 Tübingen, Germany*

ARTICLE INFO

Article history:
Accepted 18 August 2006
Available online 17 March 2009

Keywords:
Psychological refractory period
Serial and parallel processing
Optimization
Divided attention
Dual-task performance

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.

1. Introduction

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.

---

* Corresponding author. Fax: +64 3 479 8335.
E-mail addresses: miller@psy.otago.ac.nz (J. Miller), ulrich@uni-tuebingen.de (R. Ulrich).

doi:10.1016/j.cogpsych.2006.08.003

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

---

[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.