한국심리학회지 : 인지 및 생물
The Korean Journal of Cognitive and Biological Psychology
2015, Vol. 27, No. 4, 679-712

Mechanisms of capacity limits: Serial bottleneck or graded resource?

Suk Won Han†

Department of Psychology, Chungbuk National University

The current review examined how information is processed in the brain. Specifically, I investigated whether multiple inputs encoded into the brain are processed in a serial manner or they are processed in parallel. An extensive review of the literature regarding behavioral and neuroscientific studies revealed that whether information is processed in a serial or parallel manner depends on the stage of human information processing. Specifically, at the early, perceptual stage, multiple inputs can be processed in parallel as perceptual resource can be flexibly allocated to the inputs, whereas at the central stage, only a single input can be processed at a time. This review elucidates the cases in which serial or parallel processing is implemented in the brain, contributing to better understanding of how the capacity-limited brain hands information overload.

Key words : serial process, parallel process

† 교신저자 : 한석원, 충북대학교 심리학과, (28644) 충북 청주시 서원구 충대로 1
E-mail : suk.w.han@gmail.com

Introduction

The human brain is bombarded with massive amount of information, but the cognitive system cannot process all sensory inputs it receives. Strikingly, despite impressive complexity and capability of the human brain, we can barely be aware of more than a few stationary or moving objects, and we can hardly perform more than one task at the same time (Pashler, 1984; Pylyshyn & Storm, 1988; Raymond, Shapiro, & Arnell, 1992; Vogel, Luck, & Woodman, 2001).

Many daily life examples of interference between tasks or behavioral impairment have led researchers to believe that the capacity of human information processing is limited. Initially, Broadbent (1957) proposed a mechanical model to explain capacity limit. In his model, incoming information is held in a temporary stage. For this information to be used for overt response, it should pass through a kind of cognitive 'device'. He posited the presence of a bottleneck at the information processing pathway, which allows one to process only one input at a time. Thus, faced with multiple inputs, a single input should be processed completely before the next input begins to be analyzed, yielding serial processing (Pashler, 1984, 1998).

Alternatively, a resource theory was proposed by a group of researchers (Kahneman, 1973; Kok, 1997; Navon & Miller, 2002). The term 'resource' can be defined in various ways. In the present review, resources will be conceptualized as 'energetical' system that modulate or mediate certain cognitive processes (Kok, 1997). For example, the registration of incoming stimuli in visual system can be accomplished independently of attentional resource, but this perceptual process can be modulated by allocation of resource. On the other hand, there are certain processes that cannot be done without allocating atttentional resource, such as response selection or working memory consolidation and maintenance.

While Broadbent suggested that sensory inputs should pass through the attentional bottleneck for those inputs to be used for response, resource theorists posit that attentional resource should be allocated to that information for further processing. It should be noted that most resource theories also posit that attentional resource can be allocated in a graded and flexible fashion, allowing one to process multiple inputs in parallel according to task or stimulus demands (Kok, 1997; Pashler 1998). When multiple items or tasks need to be processed simultaneously, the limited resource is divided, thereby compromising processing efficiency of each task.

The current review examines whether the capacity of human information processing is limited by a serial bottleneck or the finite amount of graded resource. Before discussing mechanisms of capacity limits, a review of experimental paradigm and results demonstrating capacity limits will precede. This is necessary because a lot of experimental results believed to reveal capacity limits are susceptible to alternative accounts without assuming a capacity limit of the human cognitive system. After reviewing experimental findings for the presence of the capacity limit, results of behavioral and neuroscientific studies supporting the serial bottleneck or graded resource will be examined. Specifically, the patterns of results believed to be in favor of the serial bottleneck or limited resource will be introduced, and then, it will be critically evaluated whether the observed patterns of behaviors or neural activities can provide convincing evidence for either type of the capacity limit. To foreshadow, the review of the literature supports for the existence of both serial bottlenecks and limited resources, to account for capacity limited processing. At the end of this paper, a theoretical framework will be proposed to explain why serial processing occurs in some cases, and parallel processing is observed in others.

Demonstration of capacity limits

Does visual search reveal the capacity limit? The most common example referred to as evidence for capacity limit is the load effect observed in visual search studies (Duncan & Humphreys, 1989; Treisman & Gelade, 1980). As the number of items in visual scenes (set-size) increases, reaction times to find a specified target increases or accuracy drops. Likewise, search performance also suffers when the target and non-target items (distractor) are highly similar to each other (Duncan & Humphreys, 1989). It has been presumed that the load effect originated from the limited attentional capacity of perceptual identification of visual stimuli.

However, a number of alternatives should be considered before attributing any kind of load effect as being attentional in nature. First, the load effect observed by increasing the set-size might be due to statistical decision error, which is immune to any capacity limit (Palmer, 1994; Verghese, 2001). Assuming that sensory signals of search stimuli are noisy, the probability of confusing one of the non-target items with the target would increase as the set-size increases, regardless of any capacity limit (Huang & Pashler, 2005; Pashler, 1998). Second, when perceptual discrimination is more demanding, it has to be proven that the observed load effect was not due to any sensory factors or inherent limit in the resolution of the visual system (Lavie & DeFockert, 2003; Norman & Bobrow, 1975; Santee & Egeth, 1982). This 'data limit' is distinct from attentional capacity limit in that it exists independently of how much capacity is allocated to a given process. That is, this kind of limit cannot be resolved either by placing an input in the focus of the bottleneck or allocating more resource to that input.

To examine whether the observed load effect originated from the capacity limit instead of those confounds, Shiffrin and Gardener (1972) developed the simultaneous-sequential presentation search paradigm. Presenting items across multiple frames allows one to concentrate attentional resource on only a subset of stimuli at any instant in order to improve accuracy. The main advantage of simultaneous-sequential presentation is that this method can reveal the capacity limit without being confounded by statistical decision noise. Presenting stimuli sequentially would reduce the perceptual load of the display for a given period of time, but it would not change the total number of noise sources (i.e., the number of items) and the amount of statistical decision noise. Thus, a significant benefit in accuracy by sequential presentation would reflect the capacity limit. Some previous studies using letter stimuli, including Shiffrin and Gardener (1972), failed to report any difference between simultaneous and sequential presentation while other studies with more extended set-sizes showed significant benefit with sequential presentation (Kleiss & Lane, 1986; Prinzmetal & Banks, 1983).

Huang and Pashler (2005) also tested directly whether the set-size effect observed in various types of visual search tasks reveals capacity limit. They adopted a typical simultaneous-sequential presentation method. A standard conjunction search and a difficult feature search were not benefited from sequential presentation, casting doubt as to whether these searches really tapped into capacity-limited process. Importantly, they found a significant benefit by sequential presentation when the target and distractors shared common features but differed in spatial arrangement of those features (T vs. L). Given the significant benefit by sequential presentation in cases mentioned above, it is quite safe to conclude that the load effect observed in visual search with appropriate choice of stimuli (spatial configuration search & face search) reflects that human information processing capacity is limited (see also Han & Jung, 2 015).

Does the Psychological Refractory Period effect reveal the capacity limit?

The presence of the psychological refractory period (PRP) effect has also been interpreted that human information processing is capacity-limited. When people attempt to carry out two sensory-motor tasks presented in close temporal proximity, the response to the second task is dramatically slowed (Telford, 1931).

While the load effect in visual search reveals the capacity limit of perceptual identification of visual stimuli, numerous researches suggested that the PRP originated from the capacity limit at a central, amodal stage of information processing (McCann & Johnston, 1992; Navon & Miller, 2002; Pashler, 1984, 1994a; Tombu & Jolicoeur, 2003).

One influential model of the PRP, the 'serial bottleneck' model, proposes that dual task deficit arises from the fact that only a single response selection can proceed at a time due to the presence of a bottleneck at the response selection stage (Pashler, 1984, 1994a). Even though the serial bottleneck model has been challenged by the graded capacity sharing model suggesting parallel processing with graded resource (Tombu & Jolicoeur, 2002; 2003; 2005, Navon & Miller, 2002), these two models share a common assumption that response selection is a capacity-limited process.

Contrary to those capacity-limited views, some computational models and experimental results (Meyer & Kieras, 1997; Schumacher et al., 2001) suggest that the PRP effect was produced by participants' strategic choice to delay the second response. Specifically, Schumacher et al. (2001) claimed that dual task costs could be abolished after extensive practice during which participants were required to respond within a response time deadline. However, Tombu and Jolicoeur (2004) performed similar experiments and found a significant dual task cost even after extensive practice with more appropriately estimated baseline to compare performance between the single and the dual task condition. Furthermore, a previous study (Ruthruff, Johnston, & Remington, 2009) adopted a response time deadline procedure to encourage participants not to strategically delay the second response, yet dual task cost was unavoidable. Overall, it is quite clear that the PRP effect reveals the capacity limit of response selection even though how this dual task deficit is produced is still controversial (Han & Marois, 2013).

Does the attentional blink (AB) reveal the capacity limit? Compared to the cases of visual search and the PRP, it is less clear how the capacity limit is directly causing the attentional blink (AB). In the AB experiment, participants are searching for two targets (letters) in a rapid, serial visual presentation of distractors (digits). What is usually observed is severe impairment in reporting the second target (T2) when it was presented within approximately 500 ms from the onset of the first one (T1). To be noted, this impairment in reporting the T2 was observed only when participants were required to attend to the T1 (Raymond, Shapiro, & Arnell, 1992, for a review, see Dux & Marois, 2009).

A prominent model suggests that the main cause of the AB is the capacity-limited process of consolidating the T1 into working memory (Chun & Potter, 1995). Usually, reporting accuracy of the T2 suffers progressively as the SOA (Stimulus Onset Asynchrony) between the T1 and the T2 decreases. While the T1 is being consolidated, the T2 presented shortly after T1 does not have access to this consolidation process, and is therefore left to decay or replaced by subsequent distractors. Impaired accuracy of T2 at short SOA suggests that consolidation of perceptual input to working memory is capacity-limited (Chun & Potter, 1995; Jolicoeur, 1998, 1999).

However, there are two notable empirical observations calling for the modification of the typical capacity-limited model of the AB. As mentioned above, T2 accuracy decreases progressively as the SOA decreases. Interestingly, when the T2 follows the T1 immediately, T2 performance is relatively spared (lag 1 sparing). It is hard to explain this lag 1 sparing by T1 encoding itself because it would be presumed that the impairment would be the most severe at this shortest SOA. To explain this lag 1 sparing, the limited capacity model of the AB should be equipped with additional assumption of sluggishness of attentional selection (Di Lollo, Kawahara, Ghorashi, & Enns, 2005). Furthermore, a previous study (Vogel, Woodman, & Luck, 2006) showed that working memory consolidation per se could be accomplished within 50 ms, which was much faster than the consolidation duration estimated from the AB paradigm (about 500 ms).

Given that the AB cannot be solely explained by T1 encoding process, several groups of researchers suggested temporary loss of attentional control (Di Lollo, et al., 2005; Kawahara, Kumada, & Di Lollo, 2006), delay of attentional engagement (Nieuwenstein, Chun, van der Lubbe, & Hooge, 2005; Nieuwenstein, Potter, & Theeuwes, 2009), or overinvestment of attentional resource to T1 (Olivers & Nieuwenhuis, 2005, 2006) as primary factors for the AB instead of encoding the T1 into working memory. These studies showed that the AB could be abolished even when the number of targets to be encoded was increased as long as the targets were presented consecutively without any intervening distractor (Di Lollo et al., 2005) or blank interval (Nieuwenstein et al., 2009). More surprisingly, the AB was attenuated when a concurrent task was added while partici pants were searching for the targets in the rapid serial visual presentation (RSVP) of distractors (Olivers & Nieuwenhuis, 2005, 2006).

The fact that the AB cannot be explained solely by encoding process of T1 into working memory does not exclude attentional demand of the T1 as the factor to affect the AB. Several behavioral studies demonstrated that the AB could be obtained by increasing attentional demand of the T1, even under conditions where there used to be no AB (Dux, Asplund, & Marois, 2008, 2009). Specifically, Dux et al. replicated the pattern of results reported by Di Lollo et al. and Nieuwenstein et al., such that there was no AB when three targets were presented consecutively. They noted that T1 performance under such three-target conditions is lower than under the normal AB conditions, raising the prospect that the absence of an AB may be largely explained by increased attentional weighing of T2 at the expense of the T1. Consistent with this possibility, when the T1 required more attentional resource, a significant AB was again observed.

Serial vs. Parallel - behavioral evidence

Visual search The set-size effect in visual search has been interpreted as evidence for the capacity limit of perceptual identification of visual stimuli. One way to explain the observed set-size effect is to posit that search process is serial. Specifically, attention is focused on a single item at a time, and after processing that item, attentional focus is shifted to other item until the target is found (Pashler, 1998; Treisman & Gelade, 1980). In contrast, other researchers pointed out that the set-size effect could also be explained by limited parallel processing (Townsend, 1972, 1990). That is, the processing of one stimulus in the search display could proceed simultaneously with processing of other stimuli without waiting for completing the processing of the first stimulus. As the number of items to be processed increases, limited amount of attentional resource allocated for each item would be reduced, decreasing processing rate.

The serial search model has been prominent since Treisman and Gelade (1980) proposed the Feature Integration Theory (FIT). They reported a series of visual search experiments showing a significant set-size effect when the target was defined by a conjunction of simple features, such as color, orientation, or size. In contrast, detecting the target defined by a single feature was performed efficiently, independent of the set-size. Based upon these findings, it was argued that to integrate simple features to form a visual object, attention has to be focused on that object. To find a target defined by a conjunction of simple features, serial scanning of stimuli is necessary to focus attention on each item. Evidently, searching for the target defined by a conjunction of features is attentionally demanding, but it was not directly manifested that each item was processed serially. The mere presence of the set-size effect can also be explained by limited parallel processing (Han, 2015).

On the other hand, Duncan and Humphreys (1989) proposed a search model based upon parallel processing. However, they did not provide direct evidence supporting the claim that the set-size effect was the product of limited parallel processing, either. What they found was that target-distractor similarity and distractor-distractor similarity affected search efficiency. When target-distractor similarity was high or distractor-distractor similarity was low, a significant set size effect was observed even with a simple feature search. The steep search slope in the conjunction search could have originated from the fact that the target and distractors were made highly similar by sharing a feature. Definitely, Duncan and Humphreys' results called for modification of the FIT's main claim that conjunction search was inefficient and feature search was efficient. They suggested that feature search and conjunction search are on a continuum arguing against the claim that these two searches are performed qualitatively different ways. It is quite clear that the flat or shallow search slope observed in the feature search task is the product of parallel processing. Acknowledging that the steep search slope could be induced by the same search process underlying parallel search, Duncan and Humphreys reasoned that all searches were parallel, and the set-size effect was induced by limited parallel processing.

While neither Treisman and Gelade (1980) nor Duncan and Humphreys (1989) provided any exclusive evidence either for serial or limited parallel processing, some researchers tried to infer the mechanism of search by examining how distractors are processed when attention was directed to the target. Carrasco and Yeshurun cued the target location in the visual search task to directly examine the role of attention in search process (Carrasco & Yeshurun, 1998). Participants performed a standard conjunction visual search task. The target was presented along with a large number of distractors. On some trials, attention was oriented to the target location by a peripheral cue, and on other trials, a neutral cue was presented, providing no information about target location or identity.

Carrasco and Yeshurun (1998) reported two notable findings. First, precuing the target location diminished the eccentricity effect. Cuing enhanced behavioral performance more dramatically when the target was presented at the periphery than when it was presented close to the fovea, yielding a significant interaction between cuing and eccentricity. This suggests that the role of covert attention is to enhance sensory signal at perceptual stage of visual processing. This enhancement of sensory information by attention can fit with resource theory, given that resource allocation has been conceptualized as a sort of gain mechanism to amplify incoming signal (Kok, 1997). Second, the set-size effect was still significantly present in the cued condition, though to a lesser degree relative to the uncued condition. The strict serial model would predict that cuing would eliminate the set-size effect completely because attention was oriented to the target first, which would terminate search process (serial, self-terminating search, Treisman & Gelade, 1980). The fact that precuing enhanced performance, but did not eliminate the set-size effect implied that some irrelevant distractors were processed even when attentional weight was biased to the target, consistent with limited parallel processing.

Behavioral evidence for the presence of limited resource can also be found from Lavie and colleagues' series of experiments. They proposed a load theory of selective attention, originally intended to resolve the debate between the early and late model of attentional selection (Lavie, 1995, 1997; Lavie, 2005; Lavie, Hirst, De Fockert, & Viding, 2004). The load theory assumes that limited attentional resource is distributed across the visual scene, and each item in the scene takes up some portion of capacity in an automatic fashion. Most attentional resources would be allocated to the task relevant stimuli, and other irrelevant items would also receive attention. If target processing was relatively simple (low perceptual load), less attention is required for the target, and substantial amount of remaining attentional resource could be 'spilled over' to irrelevant visual stimuli. In this condition, irrelevant visual inputs (distractors) would be processed to such an extent that it could affect performance of the main task. In contrast, if processing of relevant visual inputs requires much attentional resource (high perceptual load), the amount of attention that irrelevant inputs receive would be reduced, leading to filtering out of irrelevant information from processing at the early stage. In this case, the effect of irrelevant items on target processing is minimized.

To test the load theory, Lavie and colleagues adopted a modified flanker task. Participants performed a letter discrimination task. One of two prespecified target letters, assigned to distinct manual responses, was presented in one of the locations on an imaginary center circle. The target could be presented by itself (low perceptual load) or with non-target items, not assigned to any response, on the center circle (high perceptual load). At the periphery, a single distractor was presented, which could be the same letter as the target (congruent), or another target that could introduce response conflict (incongruent). What was measured was the congruency effect by the flanking distractor (RT incongruent - RT congruent). Significantly slower reaction times in incongruent trials would mean that irrelevant distractors captured attention and interfered with the main task. The results showed that the flanking distractors interfered with the main task only in the low perceptual load condition. This would presumably result from the low perceptual load of the main task that did not exhaust attentional resource, leaving the remaining resource to be allocated to the distractor. In contrast, in the high load condition, there was no effect of the flaking distractor because the attentionally demanding task consumed all capacity, leaving no resource to be taken by the distractor.

The patterns of behavioral results of flanker experiments fit well with the concept of limited resource and parallel processing. However, the interpretation of the observed congruency effect is not simple. The significant congruency effect clearly means that the distractor item was processed to such an extent as to interfere with the main task, but the specific mechanism is under debate (Cho, Lien, & Proctor, 2006; Lachter, Forster, & Ruthruff, 2004). One possibility is that interference from the flanking distractor would reflect that the target and the distractor were processed in parallel with more attentional weight to the target, as Carrasco and Yeshurun argued. Alternatively, attention could be serially deployed to the target first, and after target processing, attention would be reallocated to task-irrelevant items before the response is executed.

In line with Carrasco and Yeshurun, Lavie and colleagues also inferred that parallel processing occurred based upon how the distractor was processed. If the distractor had a significant effect on the target performance, it was assumed that the distractor was processed in parallel to the target. However, the observed pattern in these studies could stem from the fact that orienting of attention was serial, but the spatial precision of the attentional focus was loose, such that some distractors were included in the attentional focus. Instead of indirect inference based upon the effect of distractor, Awh and Pashler tried to show more direct evidence for parallel processing (Awh & Pashler, 2000). They adopted a visual search paradigm in which two targets were presented in different hemifields. It was clearly shown that attentional foci could be spilt to two separate, noncontiguous locations flexibly, allowing parallel processing of each target in the opposite hemifields (see also Kawahara & Yamada, 2006).

Even though visual search has been a leading paradigm and produced fruitful outputs for study of capacity limits for a couple of decades, the interpretation of results stemming from this paradigm has never been simple (Chelazzi, 1999). However simple the stimuli and the task procedure used were, the standard search process encompasses multiple processing stages from perceptual encoding of stimuli to consolidation of perceptual information into short term memory to be used for overt responses (Duncan & Humphreys, 1989). It is also possible that both serial and parallel process exist, but at different stages of processing (Han, 2015). The problem is that there has not been agreement as to what pattern of behavioral results indicates serial or parallel processing.

PRP The current debate between the serial bottleneck and limited resource has also been extensively investigated in the context of the PRP. The most prominent model of the PRP, the 'serial bottleneck' model, suggests that dual-task limitation arises from the fact that multiple response selections can only be performed serially at a central stage of information processing (Pashler, 1984, 1994a, 1994b). This bottleneck model provides a straightforward explanation for why it is usually the second of the two responses that is slowed as the temporal interval between the two tasks (SOA) decreases, while reaction times (RT) of the first task (Task 1) remains constant (Fagot & Pashler, 1992; McCann & Johnston, 1992; Pashler & Johnston, 1989).

Despite the serial bottleneck model's parsimoniousness and predictive power, an alternative framework was proposed. Navon and Miller (2002) and Tombu and Jolicoeur (2003) developed graded capacity sharing model based upon the concept of resource that can be allocated in a graded and flexible fashion, allowing parallel processing. For example, if limited resource was evenly halved between two tasks presented at the short SOA, duration of response selection for each task would be doubled, compared to the duration of response selection at the long SOA. Viewed in this framework, serial postponement of the second response is a special case of graded sharing where the proportion of capacity allocated to the first task (sharing proportion or SP) is 100 %. With a dynamically varying SP from 50:50 to 100:0, the graded sharing model can explain every aspect of PRP data that the serial bottleneck model predicted (Tombu & Jolicoeur, 2003).

One important advantage of the graded sharing model is that it can explain the occasionally observed phenomenon of Task 1 slowing, a phenomenon that the serial bottleneck model cannot easily explain. Most PRP studies shows that the second of the two responses is slowed as SOA between two tasks decreases, while reaction times (RT) of the first task (Task 1) remain constant. However, some PRP studies have reported that decreasing SOAs not only slow down Task 2, but Task 1 as well, though to a much smaller degree than Task 2 slowing (Navon & Miller, 2002; Pashler, 1991; Sigman

& Dehaene, 2006; Tombu & Jolicoeur, 2002, 2003, 2005). An effect of SOA on Task 1 RT, however small it might be, poses a serious problem to the serial bottleneck model while the graded sharing model can explain Task 1 slowing by setting SP less than 100 %.

Interestingly, Task 1 slowing accompanied with a more severe PRP was reported by one of the studies suggesting parallel processing. Miller, Ulrich, and Rolke (2009) reasoned that participants would adopt parallel processing if it was more efficient. To create a situation where parallel processing is more efficient, the probability of SOA between two tasks was manipulated. In some blocks, short SOA trials were relatively frequent, while long SOA trials were more frequent in the other blocks. They assumed that parallel processing would be favored when short SOA trials were frequent. They further predicted that adopting parallel processing would lead to increased Task 1 RT across all SOAs (with no effect of SOA on Task 1 RT) because some portion of capacity is taken up by Task 2. The PRP deficit would be less severe because Task 2 also has access to central capacity while Task 1 is processed. Consistent with their hypothesis, Miller et al. reported that Task 1 RT was increased, and the PRP deficit was reduced, when short SOA trials were frequent. When long SOA trials were frequent, the PRP deficit was more severe as predicted. However, there was significant Task 1 slowing across all SOAs with the more frequent Long-SOA trials, a result that was not expected by Miller et al. 's original framework. As Miller et al. pointed out, the observed Task 1 slowing when long SOA trials were frequent might have been due to the fact that Task 1 performance was interfered more when short SOAs were unexpected. This delayed response to the first task at the short SOA would also lead to increased Task 2 RT, yielding steeper Task 2 RT slope (more severe PRP deficit). These results suggest that graded sharing might not be the only source of the observed Task 1 slowing.

To summarize, the presence of Task 1 slowing, which so far was believed to reflect parallel processing in the PRP, is in fact no guarantee that capacity sharing has happened. In addition, the assumption that the central capacity can be flexibly allocated across tasks has been experimentally rejected. We therefore conclude that the central stage in human information processing is limited by a serial bottleneck (see also Han & Marois, 2013).

Serial vs. Parallel – neuroscientific evidence

Neuroscientific studies can complement behavioral studies to resolve the current debate, but it has been challenging to distinguish between a serial bottleneck or graded resource because several experiments were not designed to specifically distinguish between these alternatives. Most of results reported in brain imaging studies can be intuitively linked to the concept of resource in that increasing task load is accompanied with enhancement BOLD signal. In this section, neuroscientific studies suggestive of graded resource will be selectively reviewed first, and review of data supporting serial processing follows.

Evidence for graded resource in the brain Some ERP studies have reported sensitive components to the allocation of attentional resources. One of them is a P3 component known to be sensitive to appearance of oddball or attended stimuli (Kok, 1997). In a dual-task study by Isreal and colleagues (1980), participants were primarily engaged in monitoring a visual display to detect changes in intensity or direction of squares and triangles moving on a simulated air traffic control display. The participants were also required to count the deviant stimuli in an auditory oddball task. What was measured was the amplitude of the P3 component responsive to the secondary oddball detection. As the primary task load increased (from monitoring 4 to 8 items), the amplitude of P3 was attenuated, indicating that attentional resource allocated to the secondary task was reduced. Similar results were also reported in Kramer, Wickens, and Donchin (1983). These authors required participants to control the position of the cursor with a control stick, while counting auditory or visual probes embedded in the main task. As the primary task became more difficult, the amplitude of P3 to the secondary task was reduced.

Wickens, Kramer, Vanasse, and Donchin (1983) provided further support for the presence of graded resource. They adopted a similar dual-task paradigm to previous studies, but measured P3 to both the primary and the secondary task. As the primary task load increased, the P3 to the primary task was enhanced, and the P3 to the secondary task was reduced. Along with findings showing reduced P3 to the secondary task with increased primary task load, the reciprocal tradeoff of P3 amplitude between primary and secondary tasks suggested that attentional resource could be allocated in a graded fashion.

Numerous fMRI studies also demonstrated how capacity limits are represented in the brain. One of the most common ways to investigate the capacity limit is to examine neural correlates of observed behavioral impairment as a given task load increases. Using a multiple object tracking paradigm, Culham, Cavanagh, and Kanwisher (2001) investigated which brain area is sensitive to increased load of the task. In this study, nine randomly moving objects were presented and participants tracked one to five specified targets, or simply viewed the visual display passively. Culham et al. hypothesized that tracking more items would require more attentional resource, which would increase activation in attention related brain areas. The results showed that fronto-parietal areas including Superior Frontal Sulcus, Precentral Sulcus, and Intra Parietal Sulcus showed load-dependent activation, whereas the frontal eye field (FEF) was activated by the task per se, independent of tracking load.

Instead of increasing the load of a single task, Bunge, Klingberg, Jacobsen, and Gabrieli (2000) used a dual task paradigm, and provided results that can be explained by resource theory. In this study, participants performed a single task (either reading five sentences or remembering the final words of five consecutive sentences) or both tasks in separate blocks. The results showed that performing both tasks enhanced activation of prefrontal areas involved in each single task rather than recruiting additional area that was not activated by either task. They concluded that adding additional task took up more attentional resource, which was reflected by increased BOLD signal in the task-related brain areas.

Other studies using a dual-task paradigm showed that the load dependent brain activation was observed not only in the fronto-parietal networks, but also in primary sensory cortex. Johnson and Zatorre (2005, 2006) presented visual and auditory inputs simultaneously, and participants had to attend to one sensory input selectively, or attend to both stimuli. The results showed the pattern of resource tradeoff between two sensory cortices. Attending to one sensory stimulus selectively increased activation in relevant sensory areas with decreased activation in irrelevant sensory cortices.

fMRI has allowed cognitive neuroscientists to localize areas involved in capacity limits and to get insights into where attentional resources originate. However, the caveat of fMRI research is that it cannot tell limited parallel processing from serial processing due to its poor temporal resolution. Even though many patterns of fMRI results are suggestive of the presence of limited resource that can be distributed flexibly, there has been no direct evidence to rule out the serial bottleneck exclusively. Like the cas es of behavioral studies, patterns of brain activation supporting capacity-limited parallel processing fits well with most of predictions of serial models. In the case of multiple object tracking, the specific mechanism of how people track moving objects has been under extensive debate. Specifically, people might move their focal attention rapidly from one target to the other, or spread a single attention across multiple targets (Pylyshyn & Storm, 1988; Scholl, Pylyshyn, & Feldman, 2001; Yantis, 1992). The presence of brain areas showing load-dependent activation suggests that a given area is involved in attentional process of that task, but it cannot tell how tracking is achieved. Likewise, most of the brain imaging results from dual-task paradigms presented above can also be explained by rapid switching of attention between two tasks or two distinct sensory inputs.

There are two previous studies supporting the resource theory in more convincing ways. One of them provided evidence that spatial attention can be split to two separate locations (McMains & Somers, 2004). In this study, participants had to monitor two RSVP streams of letters presented in the periphery of the visual field to detect a digit from each target stream, and tell whether the two target digits were identical. At the fovea, a digit distractor stream was presented to interfere with the main task. To perform the task, the digit distractor stream at the fovea should be inhibited effectively. The fMRI activation patterns showed that attentional foci could be split to two separate peripheral locations sparing the fovea. Increased activation was observed only in the functionally defined ROIs corresponding to the attended RSVP streams in the periphery. A plausible way to interpret McMains and Somers' results is to posit a pool of limited attentional resource that can be distributed across different locations, enabling attending multiple locations simultaneously. To be noted, it is unlikely that rapid shifting of attention could induce the observed pattern because stimulus duration per each frame of the RSVP was too brief (173 ms). This study provides converging evidence that attentional resource can be distributed over multiple disparate locations, and these multiple attentional foci can be localized in early retinotopic visual cortex.

The other study (Shim, Alvarez, Vickery, & Jiang, 2009) also provided evidence for the resource theory. A novel attentive tracking task was adopted to investigate how the number of attentional foci and the precision of each attentional focus would affect behavioral performance and brain activity. In Shim et al.'s study, participants viewed four rotating pinwheels and tracked a target spoke of one or two cued pinwheels. The number of pinwheels to be tracked and rotating speed were manipulated independently to quantify the effect of two different types of attentional demands on fronto-parietal network and perceptual areas of the brain. Behavioral performance suffered from increasing the number of pinwheels to be monitored as well as increasing the rotating speed. The fMRI results showed that two different brain networks were involved in different attentional demands. The FEF and early visual areas showed increased activation with both an increase in the number of tracked items and in the rotating speed. However, posterior parietal areas were only sensitive to the number of tracked items. Increased activation of Posterior Parietal Cortex can be explained both by rapid serial processing and parallel processing with limited resource. However, the point to be noted here is that increasing the rotating speed would not require any rapid shifting of attention. Thus, increased activation of brain areas by faster rotating speed can be interpreted as limited resources being more allocated as the task load increases.

One of the influential cognitive models based upon the concept of limited resource, the load theory by Lavie and colleague, has also been supported by numerous fMRI studies. Rees and colleagues (Rees, Frith, & Lavie, 1997) demonstrated that neural response evoked by irrelevant motion presented in the background was modulated by attentional demands of the main task presented at the fixation. Pinsk and colleagues (Pinsk, Doniger, & Kastner, 2004) and Schwartz et al. (2005) also showed a similar pattern of attentional modulation on task-irrelevant neural response occurring in V4 and V1, respectively. Furthermore, Yi and colleagues' fMRI study demonstrated that activation in the ventral visual area is also modulated by availability of attentional resource (Yi, Woodman, Widders, Marois, & Chun, 2004). While participants were performing a low attentional load task presented at fixation, repetition suppression induced by background scene images was observed in the scene-selective area (parahippocampal place area). When the main task became more perceptually challenging, the amount of attention spilled over to irrelevant background scene was reduced, and repetition suppression was abolished.

Some of the ERP and fMRI studies reviewed above showed, equivocal evidence in favor of the presence of limited resources that could be distributed in a graded and flexible fashion. The point to be noted is that experimental paradigms used in those studies require participants to continuously maintain attention to spatial locations, moving objects, or task sets. The observed pattern of brain activity that can be linked to limited resource might have stemmed from the fact that those particular paradigms require subjects to activate a single or multiple attentional sets in a sustained manner.

Evidence for serial processing in the brain Given behavioral evidence that the PRP is the product of the serial bottleneck at response selection stage, it is useful to examine neural activity under dual task situation to observe serial processing in the brain. Jiang, Saxe, and Kanwisher (2004) performed an fMRI experiment to investigate the mechanism of the PRP. They presented two visuo-manual tasks either in a short SOA or in a long SOA. In the short SOA, the response of the second task was severely slowed (the PRP effect), but any brain area related with the behavioral dual task deficit was not found. Based upon lack of increased activation in any brain area under dual task situation, Jiang et al. concluded that multiple response selection is performed by passive, serial delay of the second response without any executive control.

The limited temporal resolution of fMRI makes it difficult to apply this technique to investigate the PRP because the PRP reveals a temporal limitation in performing two concurrent tasks. Considering that the PRP reveals a temporal limitation, Dux and colleagues used time-resolved fMRI (Goebel, Roebroeck, Kim, & Formisano, 2003) and measured duration of neural activity to isolate the locus of the response selection bottleneck (Dux, Ivanoff, Asplund, & Marois, 2006; Dux et al., 2009). To improve fMRI 's temporal resolution, they sampled brain activity rapidly (short TR) and employed 2 eight alternative forced choice (8 AFC) tasks to have substantial amount of dual-task costs to be captured by the BOLD response. Dux et al. measured the duration of neural activity estimated by the peak latency of hemodynamic response (Henson, Price, Rugg, Turner, & Friston, 2002). Participants performed visuo-vocal and auditory manual 8 AFC tasks presented either with a short or a long SOA. Behavioral data in the scanner showed a robust PRP, and this PRP effect was related with prolonged duration of BOLD response in the left Infeior Frontal Junction area. Specifically, at the short SOA, duration of BOLD activity was prolonged for slow Task 1 RTs relative to fast Task 1 RTs, but at the long SOA, there was no difference in BOLD activity between slow RTs and fast RTs. At the short SOA, slow Task 1 RT prolonged the duration of overall response selection process because of temporal overlapping of two response selections. At the long SOA, however, slow Task 1 RT does not affect the total duration of response selection process because those response selection processes were temporally separate. The results of Dux et al. suggests that duration of neural activity in the inferior frontal junction(IFJ) is sensitive to duration of response selection process, implying this area is the neural locus of response selection bottleneck.

Even though Dux et al. (2006; 2009) provided convincing evidence that the IFJ was a neural locus of the central bottleneck, the pattern of neural activity in this area related with dual task deficit could not distinguish between the serial bottleneck and graded capacity sharing models as the two models have identical prediction for behavioral and neural data.

The fact that evidence for serial processing came from brain imaging studies of the PRP (Dux et al., 2006; 2009; see also Sigman & Dehaene, 2008) does not mean that information is processed serially only at the amodal, central processing stage. Indeed, the most straightforward and direct evidence for serial processing in the brain can be found from a couple of visual search studies, requiring participants to perform perceptual identification of visual inputs. Woodman and Luck (1999, 2003) measured electrophysiological responses while human participants were performing a visual search task. High temporal resolution of the ERP technique allowed them to measure the moment-by-moment allocation of attention to distinguish whether search process was serial or parallel. As Woodman and Luck correctly pointed out, the appropriate choice of search task was crucial to investigate how search proceeds because the load effect does not guarantee that a given search task requires limited resource. To avoid any potential confound by decision noise or data limit, they used a visual search task in which the target and distractors shared common features and differed only in spatial arrangement of those features. Specifically, the target was a square that had a gap on a particular side, and distractors were squares that had a gap on a different side.

Woodman and Luck (1999) presented four different colored items in a separate quadrant along with black distractors. Among four colored items, a target was presented in a prespecified color (C75) on 75 % of trials, and in a different color (C25) on remaining 25 % of trials to bias participants to prioritize items in C75. Presumably, participants would prioritize items in C75 and C25 over the items in other colors. Furthermore, in most trials, the target was in C75, thus, the item in C75 would be further prioritized over the item in C25. The way to test whether the search was serial or parallel is to present the items in C75 and C25 in the opposite hemifields and examine if there is any interhemishperic shift of focal attention. The N2pc component of the ERP waveforms, which is believed to index covert orienting of spatial attention, was measured while participants were performing visual search task. What these studies found was rapid shifting of the N2pc component across hemifields during search process, as predicted b y the serial search model.

Buschman and Miller (2009) also provided behavioral and neurophysiological evidence for serial search. They trained monkeys to perform an effortful search (top-down) yielding a steep search slope and a pop-out search in which the search slope was almost flat. Contrary to

Woodman and Luck who explicitly biased participants to search stimuli in a given order, Buschman and Miller did not bias monkeys to adopt any search strategy. Behavioral results suggested that monkeys happened to adopt the strategy to scan each item in a clockwise manner until the target was found during the top-down search task, whereas no ordered pattern was found during the pop-out search. Across all recording sessions, search reaction times were shortest when the target was in the lower-right position and became progressively longer as the target appeared in the lower-left, upper-left, and the upper-right position. Serial search patterns observed in behavioral results were also reflected in neuronal activities in the FEF. When the target appeared at the location clockwise from the neuron's preferred location, a transient increase of FEF activity was observed before the target was found, which suggested that the focal attention shifted in a clockwise manner from the neuron's preferred location to the target location.

Discussion

From the review so far, I have drawn the conclusion that both serial and parallel processes with limited resource may exist in the human brain. However, it has yet to be specified why serial processing may occur in some cases, and parallel processing in others. In this section, a framework that can encompass both serial and limited parallel processing is introduced, and new empirical evidence supporting that framework is presented.

Attention-Perceptual vs. Central One possible solution to reconcile serial and parallel processing is to set separate capacity limits of attention for different processing stages and to posit that attention works differently at each processing stage. Pashler and colleagues argued that there existed separate and distinct capacities for perceptual and central attention (Pashler, 1991; Pashler & Johnston, 1989). In a series of experiments in Pashler (1991), participants were required to perform a tone discrimination task and identify visual stimuli presented briefly following the tone. Requiring participants to immediately respond to a simple sound did not impair perceptual identification of visual stimuli. Considering that multiple response selections are performed serially (Pashler, 1984, 1994a, 1994b) while spatial attention can be split to two separate locations (Awh & Pashler, 2000), Pashler suggested that central attention works serially whereas perceptual attention operates in parallel.

Contrary to previous studies suggesting that central and visuospatial attention are independent (Johnston, McCann, & Remington, 1995; Pashler, 1991), more recent studies indicate that these two types of attention may interact. A group of researchers demonstrated that loading central attention interfered with covert orienting of spatial attention (Brisson & Jolicoeur, 2007). In Brisson and Jolicoeur's electrophysiological studies, a similar version of the PRP paradigm that Pashler and colleagues used was adopted. The first task was to make an immediate and speeded response to a tone presented briefly. The second task, a visual search task requiring shifting of spatial attention, followed either at short SOA or long SOA. While participants were performing the task, the N2pc component indexing covert orienting of spatial attention (Woodman and Luck, 1999, 2003) was measured. The N2pc amplitude in the short SOA condition was reduced, meaning that loading central attention by requirement of an immediate response to the simple tone interfered with deployment of visuospatial attention to the search target. Jolicoeur and colleagues suggested that perceptual attention and central attention interact.

As Pashler (1998) pointed out, it is highly misleading to understand attention as a unified and homogenous process, especially given that attentional selection is serial in some cases, and parallel processing is observed in others. However, distinguishing between perceptual and central attention cannot explain more recent behavioral and neuroscientific findings that shows the interaction between perceptual and central attention. Moreover, the fact that allocation of visuospatial attention can be accomplished serially (Buschman & Miller, 2009; Woodman & Luck, 1999, 2003) challenges the claim that serial processing occurs only at the central stage where response selection or decision-making is made.

Attention–Control vs. Resource Considering that the typical view of distinct capacity limits for different processing stages cannot accommodate recent empirical evidence, an alternative framework to explain the presence of both serial and parallel processing need to be proposed. The present review of behavioral and neuroscientific studies showed that the pattern of serial processing was observed when people are required to make immediate response selections, or to switch attention between distinct cognitive sets or separate locations. On the other hand, parallel processing with graded resource seems to occur when sustaining attention to multiple locations, moving objects, or cognitive sets is necessary.

Based upon the trends revealed from the current review so far, it is suggested that attention can be decomposed into a control process that works transiently, and a pool of limited resource that is manifested when attention has to work in a sustained manner.

Attentional control plays a critical role to adjust behavior in a rapid and transient way to a dynamically changing environment. Specifically, this control process is crucial to switch between different cognitive sets, to shift attention from one location to another, or moment-to-moment tracking of moving objects (Tombu & Seiffert, 2008). One important point to be mentioned is that it exerts only a single operation at a time, yielding serial processing (Di Lollo et al., 2005). On the other hand, graded resource is recruited to maintain attention to multiple locations, moving objects, and other various cognitive sets simultaneously in a sustained manner. One well-known effect of this sustained attention is enhancement of sensory information of attended objects or of objects at attended locations (Carrasco, Penpeci-Talgar, & Eckstein, 2000; Carrasco, Williams, & Yeshurun, 2002; Carrasco & Yeshurun, 1998; O'Craven, Downing, & Kanwisher, 1999; Reynolds & Heeger, 2009). Enhancing the strength of perceptual representations can be accomplished via distractor suppression or signal enhancement (Awh, Matsukura, & Serences, 2003; Awh, Sgarlata, & Kliestik, 2005; Dosher & Lu, 2000; Lu & Dosher, 1998; Lu, Lesmes, & Dosher, 2002; Prinzmetal, McCool, & Park, 2005; Serences, Yantis, Culberson, & Awh, 2004), and all these processes are resource-consuming (Engle, Conway, Tuholski, & Shisler, 1995).

Decomposing cognitive process into transient and sustained component is not entirely novel. A previous fMRI study also proposed that cognitive control could be dissected into transient and sustained processes (Braver, Reynolds, & Donaldson, 2003). In this hybrid event-related and blocked fMRI experiment, participants were exposed to two types of blocks. One was a single-task block, in which participants were performing a single, homogenous task (pure block) throughout. In the other block (mixed block), two different tasks were intermixed, and for each trial, participants were instructed which task they have to perform, hence trials could be separated into task-repeat trials and task-switch trials. Contrasting between single and mixed blocks showed that right anterior PFC was involved in non-trial specific, and sustained control, presumably related with the active continuous maintenance of multiple task sets. It was also found that the left lateral prefrontal area was related with trial-by-trial fluctuation of switching cost, suggesting that this area was exerting transient control to adjust behavior in the task switching condition. Notably, the area localized to be involved in the transient control process, lateral prefrontal cortex, is also reported to be the locus of response selection bottleneck responsible for the PRP (Dux et al., 2006; 2009). The fact that the lateral prefrontal cortex, IFJ, was crucial for both the transient control process and response selection bottleneck provides converging evidence that the area for transient control is the place where multiple inputs are processed serially.

Prinzmetal and colleagues also proposed two different types of attentional effect on behavioral performance based on a series of cuing experiments (Prinzmetal, McCool, & Park, 2005; Prinzmetal, Park, & Garrett, 2005; Prinzmetal, Zvinyatskovskiy, Gutierrez, & Dilem, 2009). First, attention can be thought of as resource allocation to increase the quality of sensory information on attended location (Channel Enhancement). Attentional effect measured by accuracy without confound of speeded response was presumed to reveal the enhancement of perceptual representation. Resource allocation is necessary to observe this enhancement. Second, attention can select one aspect of incoming information over another. In this case, resource allocation is not necessary, and attention prioritizes one over the other without any sensory modulation (Channel Selection). This prioritization is revealed by attentional effect measured by reaction time. Thus,attentional benefit in reaction time does not require any resource allocation.

A previous behavioral study provided converging evidence for Prinzmetal and colleagues ’ distinction between attentional control and resource (Han & Kim, 2008). Han and Kim showed that attentional selection and signal enhancement via resource allocation interacted with spatial working memory (the maintenance of spatial attention on multiple locations) differently. Using a spatial cuing paradigm combined with a working memory task, they showed that spatial working memory load did not affect the cuing effect measured by reaction time, which they interpreted as indicating that attentional selection per se was intact. In contrast, cuing effect measured by accuracy in data-limited condition, which was presumed to reflect enhancement of sensory signal, was reduced by spatial working memory load. Han and Kim concluded that the maintenance of spatial attention on multiple locations interfered with resource-consuming attentional enhancement of perceptual representation.

By separating control processes from attentional resources, several observations supporting serial and limited parallel processing can be explained in an integrative way. Serial processing was observed in the PRP, which requires high demand of attentional control for immediate response selection and rapid switching of cognitive set. Deployment of visuospatial attention was also accomplished serially in the case that several distractors need to be filtered out by attentional control. Indeed, Di Lollo, et al. (2005) assumed that a control processor could exert appropriate control only once at a time. On the other hand, maintaining attention on spatially separate locations or multiple task sets can be done in a parallel and sustained manner. Furthermore, the interaction between central attention and perceptual attention can also be explained by suggesting that control process is domain or task independent (Chiu & Yantis, 2009; Duncan, 2001; Duncan & Owen, 2000).

Interaction between attentional control and resource The fact that attentional control and attentional resource can be dissociated neurally and behaviorally does not exclude the possibility of an interaction between these two components. One example is the AB. Chun and Potter (1995) suggested that capacity limited process of consolidating T1 into visual working memory left perceptual representation of T2 susceptible to decay. This limited capacity model of the AB was challenged by the Temporary Loss of Control account (Di Lollo et al., 2005; Kawahara et al., 2006). According to the TLC account, the AB is observed mainly due to the distractor immediately following the T1. What is crucial for accurate target report in a RSVP is the establishment of an attentional set to select the target(s) and filter out distractors. Di Lollo et al. showed that no AB was observed if targets were presented in succession with no distractors between those targets (uniform condition). The AB was observed only when a distractor was inserted between the targets (varied condition), disrupting attentional setting. Nieuwenstein et al. argued against Di Lollo et al. 's claim that the distractor following the T1 was the main reason for the AB because they could obtain the AB with a blank interval inserted between targets without any distractor after the T1 (Nieuwenstein, Potter, & Theeuwes, 2009). However, their results can be consistent with the claim that the AB is a matter of attentional control, given that the blank interval would introduce discontinuity in target stream and disrupt attentional setting, just as the distractor would. More serious challenges to the TLC account can be found from Dux et al. (2008, 2009). They provided results in favor of the limited capacity model of the AB because such AB could be observed even when targets were presented continuously without any intervening distractor simply by increasing the attentional demands to T1.

How can two opposing models of the AB (limited capacity vs. attentional control) be reconciled? It is evident that attentional control is a crucial factor to induce the AB as the TLC account suggests. It should also be admitted that resource allocation plays a significant role (Chun & Potter, 1995; Dux et al., 2008, 2009; Dux & Marois, 2009; Jolicoeur, 1998, 1999). Allocating more resource to the first target would make control process to switch to the next target more challenging. In the uniform condition, where no AB is observed, less control is needed relative to the varied condition. However, when required to allocate more attentional resource to the first target, all targets are not in the same attentional state any more because one target was attentionally weighted. To switch from a weighted target to unweighted one, control process would be required, as it is for the varied condition.

The neural locus of a central bottleneck and attentional control The lateral prefrontal cortex, particularly the IFJ, has been known to be involved in various types of cognitive control (Brass, Derrfuss, Forstmann, & von Cramon, 2005; Braver, Reynolds, & Donaldson, 2003). Notably, this region was also found to be the locus of the response selection bottleneck in the PRP (Dux et al., 2006; 2009). Given that the PRP is induced by the presence of a serial bottleneck (Pashler 1994; Han and Marois, 2013), the IFJ would be a brain substrate where multiple inputs are processed serially (Dux et al., 2006; 2009). Furthermore, Todd and colleagues reported that this area was involved in visual short term memory encoding (Todd, Han, Harrison, & Marois, 2011). Using a time-resolved fMRI, they showed that neural activity in this area was sensitive to the duration of encoding (consolidating) visual information into working memory. Moreover, the fact that the IFJ was involved in the AB, working memory encoding, and the PRP supports the claim that the AB and the PRP has a common origin and working memory consolidation plays a critical role to induce both the AB and the PRP (Jolicoeur, 1998, 1999; Jolicoeur & Dell'Acqua, 1998).

The IFJ is not only activated in the case that attentional control has to be applied endogenously or voluntarily. A recent fMRI study by Asplund and colleagues presented results that this area participates in stimulus-driven attention as well as goal-directed attention (Asplund, Todd, Snyder, & Marois, 2010). Moreover, from a functional connectivity analysis, it was found that this area was functionally coupled with posterior attention networks, depending on the type of attentional processing (Frontal Eye Field, Intra Parietal Sulcus for goal-directed attention and Temporal Parietal Junction for stimulus-driven attention).

All these results suggest that the IFJ is a neural locus for a common processing bottleneck from perception to action (Marois & Ivanoff, 2005), and a key area for various types of attentional control in general (Brass, Derrfuss, Forstmann, & von Cramon, 2005).

Conclusions It has been a long-standing issue in the field of cognitive psychology and neuroscience whether the capacity of human information processing is limited by a serial bottleneck or finite amount of resource. Most of data supporting each type of processing limit was usually compatible to the opposing accounts. It is also true that the current technologies have several limitations to reveal directly serial bottlenecks or limited resources. If so, is the distinction between the serial bottleneck and limited resource really impossible to prove or disprove? I do not think that is the case. Even though the review of behavioral studies of visual search could not provide a conclusive answer to the nature of the capacity limit revealed by this paradigm, neuroscientific studies contributed to resolve the issue. From the ERP and cell recording studies, it was clearly evident that serial processing occurred during visual search. In the case of the PRP, the review of behavioral studies provided quite convincing evidence that central processing (response selection) is limited by a serial bottleneck, which was hard to tell from any brain imaging method. Given that serial processing exists both at the perceptual and at the central stages, it can be suggested that serial processing occurs when rapid switching or control of attention is necessary, independently of tasks and processing stages. On the other hand, empirical supports for limited resource in the brain could be found from numerous brain imaging studies. One common element in these studies was that the experimental paradigms require sustained attention on multiple locations, moving objects, or task sets.

To explain reviewed data supporting the serial bottleneck and limited resource models, I proposed a framework suggesting the separation between attentional control operating serially (and transiently), and a pool of limited resource being allocated in a graded (also sustained) fashion. Recent fMRI studies of the PRP and the AB suggested that the IFJ could be the locus of attentional control and a common processing bottleneck from perception to action that may subserve both serial and sustained attention.

The purpose of this review was to get a better understanding of how the capacity of information processing is limited in the context of the debate between the serial bottleneck and graded resource allowing parallel processing. Due to the current absence of methodology directly telling serial process from limited parallel process, this issue has slowly been abandoned (Thornton & Gilden, 2007). However, this issue has been extensively investigated throughout the history of cognitive psychology and neuroscience, and the time to harvest the fruit of advancement in neuroscientific techniques and in the sophistication of behavioral paradigms may be upon us.

References

Asplund, C. L., Todd, J. J., Snyder, A. P., & Marois, R. (2010). A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. *Nature Neuroscience, 13*, 507-513.

Awh, E., Matsukura, M., & Serences, J. T. (2003). Top-down control over biased competition during covert spatial orienting. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 52-63.

Awh, E., & Pashler, H. (2000). Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 834-846.

Awh, E., Sgarlata, A. M., & Kliestik, J. (2005). Resolving visual interference during covert spatial orienting: online attentional control through static records of prior visual experience. *Journal of Experimental Psychology: General, 134*, 192-206.

Brass, M., Derrfuss, J., Forstmann, B., & von Cramon, D. Y. (2005). The role of the inferior frontal junction area in cognitive control. *Trends in Cognitive Sciences, 9*, 314-316.

Braver, T. S., Reynolds, J. R., & Donaldson, D. I. (2003). Neural mechanisms of transient and sustained cognitive control during task switching. *Neuron, 39*, 713-726.

Brisson, B., & Jolicoeur, P. (2007). Electrophysiological evidence of central interference in the control of visuospatial attention. *Psychonomic Bulletin & Review, 14*, 126-132.

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review, 64*, 205-215.

Bunge, S. A., Klingberg, T., Jacobsen, R. B., & Gabrieli, J. D. (2000). A resource model of the neural basis of executive working memory. *Proceedings of the National Academy Sciences of the United States of America, 97*, 3573-3578.

Buschman, T. J., & Miller, E. K. (2009). Serial, covert shifts of attention during visual search are reflected by the frontal eye fields and correlated with population oscillations. *Neuron, 63*, 386-396.

Carrasco, M., Penpeci-Talgar, C., & Eckstein, M. (2000). Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. *Vision Research, 40*, 1203-1215.

Carrasco, M., Williams, P. E., & Yeshurun, Y. (2002). Covert attention increases spatial resolution with or without masks: Support for signal enhancement [Electronic Version]. *Journal of Vision*, 2, 467-479.

Carrasco, M., & Yeshurun, Y. (1998). The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 673-692.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention.

Trends in Cognitive Sciences, 9, 349-354.

Chelazzi, L. (1999). Serial attention mechanisms in visual search: A critical look at the evidence. *Psychological Research, 62*, 195-219.

Chiu, Y. C., & Yantis, S. (2009). A domain-independent source of cognitive control for task sets: shifting spatial attention and switching categorization rules. *Journal of Neuroscience, 29*, 3930-3938.

Cho, Y. S., Lien, M. C., & Proctor, R. W. (2006). Stroop dilution depends on the nature of the color carrier but not on its location. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 826-839.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 109-127.

Culham, J. C., Cavanagh, P., & Kanwisher, N. G. (2001). Attention response functions: characterizing brain areas using fMRI activation during parametric variations of attentional load. *Neuron, 32*, 737-745.

Di Lollo, V., Kawahara, J. -I., Ghorashi, S. M. S., & Enns, J. T. (2005). The attentional blink: Resource depletion or temporary loss of control? *Psychological Research, 69*, 191-200.

Dosher, B. A., & Lu, Z. L. (2000). Noise exclusion in spatial attention. *Psychological Science, 11*, 139-146.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Review Neuroscience, 2*, 820-829.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*, 433-458.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neuroscience, 23*, 475-483.

Dux, P. E., Asplund, C. L., & Marois, R. (2008). An attentional blink for sequentially presented targets: Evidence in favor of resource depletion accounts. *Psychonomic Bulletin & Review, 15*, 809-813.

Dux, P. E., Asplund, C. L., & Marois, R. (2009). Both exogenous and endogenous target salience manipulations support resource depletion accounts of the attentional blink: A reply to Olivers, Spalek, Kawahara & Di Lollo. *Psychonomic Bulletin & Review, 16*, 219-224.

Dux, P. E., Ivanoff, J., Asplund, C. L., & Marois, R. (2006). Isolation of a central bottleneck of information processing with time-resolved fMRI. *Neuron, 52*, 1109-1120.

Dux, P. E., & Marois, R. (2009). The attentional blink: A review of data and theory. *Attention, Perception, & Psychophysics, 71*, 1683-1700.

Dux, P. E., Tombu, M., Harrison, S., Rogers, B. P., Tong, F., & Marois, R. (2009). Training improves multitasking performance by increasing the speed of information processing in human prefrontal cortex. *Neuron, 63*, 127-138.

Engle, R. W., Conway, A. R. A., Tuholski, S. W., & Shisler, R. J. (1995). A resource account of inhibition. *Psychological Science, 6*, 122-125.

Fagot, C., & Pashler, H. (1992). Making two responses to a single object: implications for the central attentional bottleneck. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 1058-1079.

Fougnie, D., & Marois, R. (2006). Distinct capacity limits for attention and working memory: Evidence from attentive tracking and visual working memory paradigms. *Psychological Science, 17*, 526-534.

Goebel, R., Roebroeck, A., Kim, D. S., & Formisano, E. (2003). Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. *Magnetic Resonance Imaging, 21*, 1251-1261.

Han, S. W. (2015). Distinct load effects by set-size and target-distractor similarity in visual search. 한국심리학회지: 인지 및 생물, 27(3), 505-518.

Han, S. W., & Jung, W. H. (2015). Your own face is no more precious than others': Evidence from the simultaneous-sequential paradigm. *Psychonomic Bulletin & Review, 22*, 1-6.

Han, S. W., & Kim, M. S. (2008). Spatial working memory load impairs signal enhancement, but not attentional orienting. *Perception & Psychophysics, 70*, 916-923.

Han, S. W., & Marois, R. (2013). The source of dual-task limitations: Serial or parallel processing of multiple response selections? *Attention, Perception & Psychophysics, 75*, 1395-1405.

Henson, R. N., Price, C. J., Rugg, M. D., Turner, R., & Friston, K. J. (2002). Detecting latency differences in event-related BOLD responses: application to words versus nonwords and initial versus repeated face presentations. *NeuroImage, 15*, 83-97.

Huang, L., & Pashler, H. (2005). Attention capacity and task difficulty in visual search. *Cognition, 94*, B101-B111.

Isreal, J. B., Chesney, G. L., Wickens, C. D., & Donchin, E. (1980). P300 and tracking difficulty: evidence for multiple resources in dual-task performance. *Psychophysiology, 17*, 259-273.

Jiang, Y., Saxe, R., & Kanwisher, N. (2004). Functional magnetic resonance imaging provides new constraints on theories of the psychological refractory period. *Psychological Science, 15*, 390-396.

Johnson, J. A., & Zatorre, R. J. (2005). Attention to simultaneous unrelated auditory and visual events: behavioral and neural correlates. *Cerebral Cortex, 15*(10), 1609- 1620.

Johnson, J. A., & Zatorre, R. J. (2006). Neural substrates for dividing and focusing attention between simultaneous auditory and visual events. *NeuroImage, 31*, 1673-1681.

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

Jolicoeur, P. (1998). Modulation of the attentional blink by on-line response selection: Evidence from speeded and unspeeded Task1 decisions. *Memory & Cognition, 26*, 1014-1032.

Jolicoeur, P. (1999). Concurrent response-selection demands modulate the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1097-1113.

Jolicoeur, P., & Dell'Acqua, R. (1998). The demonstration of short-term consolidation. *Cognitive Psychology, 36*, 138-202.

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

Kawahara, J.-I., Kumada, T., & Di Lollo, V. (2006). The attentional blink is governed by a temporary loss of control. *Psychonomic Bulletin & Review, 13*, 886-890.

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

Kok, A. (1997). Event-related-potential (ERP) reflections of mental resources: a review and synthesis. *Biological Psychology, 45*, 19-56.

Kramer, A. F., Wickens, C. D., & Donchin, E. (1983). An analysis of the processing requirements of a complex perceptual-motor task. *Human Factors, 25*, 597-621.

Lachter, J., Forster, K. I., & Ruthruff, E. (2004). Forty-five years after Broadbent (1958): Still no identification without attention. *Psychological Review, 111*, 880-913.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 451-468.

Lavie, N. (1997). Visual feature integration and focused attention: response competition from multiple distractor features. *Perception & Psychophysics, 59*, 543-556.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences, 9*, 74-82.

Lavie, N., & DeFockert, J. W. (2003). Contrasting effects of sensory limits and capacity limits in visual selective attention. *Perception & Psychophysics, 65*, 202-212.

Lavie, N., Hirst, A., De Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General, 133*, 339-354.

Lu, Z., -L., & Dosher, B. (1998). External noise distinguishes attention mechanisms. *Vision Research, 38*, 1183-1198.

Lu, Z. L., Lesmes, L. A., & Dosher, B. A. (2002). Spatial attention excludes external noise at the target location. *Journal of Vision, 2*, 312-323.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279-281.

Marois, R., Chun, M. M., & Gore, J. C. (2000). Neural correlates of the attentional blink. *Neuron, 28*, 299-308.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences, 9*, 296-305.

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

McMains, S. A., & Somers, D. C. (2004). Multiple spotlights of attentional selection in human visual cortex. *Neuron, 42*, 677-686.

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

Miller, J., Ulrich, R., & Rolke, B. (2009). On the optimality of serial and parallel processing in the psychological refractory period paradigm: Effects of the distribution of stimulus onset asynchronies. *Cognitive Psychology, 58*, 273-310.

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

Nieuwenstein, M. R., Chun, M. M., van der Lubbe, R. H., & Hooge, I. T. (2005). Delayed attentional engagement in the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 1463-1475.

Nieuwenstein, M. R., Potter, M. C., & Theeuwes, J. (2009). Unmasking the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 159-169.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 4*, 44-64.

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature, 401*, 584-587.

Olivers, C. N., & Nieuwenhuis, S. (2005). The beneficial effect of concurrent task-irrelevant mental activity on temporal attention. *Psychological Science, 16*, 265-269.

Olivers, C. N., & Nieuwenhuis, S. (2006). The beneficial effects of additional task load, positive affect, and instruction on the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 364-379.

Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision Research, 13*, 1703-1721.

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

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

Pashler, H. (1994a). Dual-task interference in simple tasks: Data and theory. *Psychological*

Bulletin, 116(2), 220-244.

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

Pashler, H. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.

Pashler, H., & Johnston, J. C. (1989). Chronometric evidence for central postponement in temporally overlapping tasks. *Quarterly Journal of Experimental Psychology, 41A*, 161-191.

Pinsk, M. A., Doniger, G. M., & Kastner, S. (2004). Push-pull mechanism of selective attention in human extrastriate cortex. *Journal of Neurophysiology, 92*, 622-629.

Prinzmetal, W., & Banks, W. P. (1983). Perceptual capacity limits in visual detection and search. *Bulletion of the Psychonomic Society, 21*, 263-266.

Prinzmetal, W., McCool, C., & Park, S. (2005). Attention: Reaction time and accuracy reveals different mechanisms. *Journal of Experimental Psychology: General, 134*, 73-92.

Prinzmetal, W., Park, S., & Garrett, R. (2005). Involuntary attention and identification accuracy. *Perception & Psychophysics, 67*, 1344-1353.

Prinzmetal, W., Zvinyatskovskiy, A., Gutierrez, P., & Dilem, L. (2009). Voluntary and involuntary attention have different consequences: The effect of perceptual difficulty. *Quarterly Journal of Experimental Psychology, 62*, 352-369.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision, 3*, 179-197.

Raymond, J., E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance, 18*, 849-860.

Rees, G., Frith, C. D., & Lavie, N. (1997). Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science, 278*, 1616-1619.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron, 61*, 168-185.

Ruthruff, E., Johnston, J. C., & Remington, R. W. (2009). How strategic is the central bottleneck: Can it be overcome by trying harder? *Journal of Experimental Psychology: Human Perception and Performance, 35*, 1368-1384.

Santee, J. L., & Egeth, H. E. (1982). Do reaction time and accuracy measure the same aspect of letter recognition? *Journal of Experimental Psychology: Human Perception and Performance, 8*, 489-501.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition, 80*, 159-177.

Schumacher, E., Seymour, T., Glass, J., Fencsik, D., Lauber, E., Kieras, D., et al. (2001).

Virtually perfect time sharing in dual-task performance: Uncorking the central cognitive bottleneck. *Psychological Science, 12*, 101-108.

Schwartz, S., Vuilleumier, P., Hutton, C., Maravita, A., Dolan, R. J., & Driver, J. (2005). Attentional load and sensory competition in human vision: Modulation of fMRI responses by load at fixation during task-irrelevant stimulation in the peripheral visual field. *Cerebral Cortex, 15*, 770-786.

Serences, J. T., Yantis, S., Culberson, A., & Awh, E. (2004). Preparatory activity in visual cortex indexes distractor suppression during covert spatial orienting. *Journal of Neurophysiology, 92*, 3538-2545.

Shiffrin, R. M., & Gardner, G. T. (1972). Visual processing capacity and attentional control. *Journal of Experimental Psychology, 93*, 72-82.

Shim, W. M., Alvarez, G. A., Vickery, T. J., & Jiang, Y. V. (2009). The Number of Attentional Foci and Their Precision Are Dissociated in the Posterior Parietal Cortex. *Cerebral Cortex*.

Sigman, M., & Dehaene, S. (2006). Dynamics of the central bottleneck: dual-task and task uncertainty [Electronic Version]. *PLoS Biology*, 4, e220. Retrieved Jul 2008 from http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0040220.

Sigman, M., & Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during dual-task performance. *Journal of Neuroscience, 28*, 7585-7598.

Telford, C. (1931). The refractory phase of voluntary and associative response. *Journal of Experimental Psychology, 14*, 1-35.

Thornton, T. L., & Gilden, D. L. (2007). Parallel and serial processes in visual search. *Psychological Review, 114*, 71-103.

Todd, J. J., Han, S. W., Harrison, S., & Marois, R. (2011). The neural correlates of visual working memory consolidation: A time-resolved fMRI study. *Neuropsychologia, 49,* 1527-1536.

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

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

Tombu, M., & Jolicoeur, P. (2004). Virtually no evidence for virtually perfect time-sharing. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 795-810.

Tombu, M., & Jolicoeur, P. (2005). Testing the predictions of the central capacity sharing model. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 790-802.

Tombu, M., & Seiffert, A. E. (2008). Attentional costs in multiple-object tracking. *Cognition, 108*, 1-25.

Townsend, J. T. (1972). Some results concerning the identifiability of parallel and serial processes. *British Journal of Mathmatical and Statistical Psychology, 25*, 168-199.

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but the can (and should) be distinguished. *Psycholgical Science, 1*, 46-54.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron, 31*, 523-535.

Vogel, E. K., Luck, S. J., & Woodman, G. F. (2001). Storage of features, conjunctions and object in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 92-114.

Wickens, C., Kramer, A., Vanasse, L., & Donchin, E. (1983). Performance of concurrent tasks: A psychophysiological analysis of the reciprocity of information- processing resources. *Science, 221*, 1080-1082.

Woodman, G. F., & Luck, S. J. (1999). Electrophysiological measurement of rapid shifts of attention during visual search. *Nature, 400*, 867-869.

Woodman, G. F., & Luck, S. J. (2003). Serial deployment of attention during visual search. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 121-138.

Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology, 24*, 295-340.

Yi, D. J., Woodman, G. F., Widders, D., Marois, R., & Chun, M. M. (2004). Neural fate of ignored stimuli: dissociable effects of perceptual and working memory load. *Nature Neuroscience, 7*, 992-996.

1 차원고접수 : 2015. 06. 21
수정원고접수 : 2015. 10. 14
최종게재결정 : 2015. 10. 19

한국심리학회지 : 인지 및 생물
The Korean Journal of Cognitive and Biological Psychology
2015, Vol. 27, No. 4, 679-712

용량제한의 기전: 순차 병목 혹은 한정된 자원?

한 석 원

충북대학교 심리학과

본 연구에서는 인간의 정보처리의 속성에 대해 고찰하였다. 외부로부터 다수의 자극 제시될 때 뇌에서 처리될 수 있는 양은 한정되어 있기에 이러한 용량제한이 어떻게 해소되는지를 이해하는 것은 인지심리학 및 인지신경과학의 중요한 화두였다. 구체적으로, 뇌에 입력된 정보들은 한 번에 하나씩, 순차적으로, 혹은 한 번에 여러 자극이 병렬적으로 처리가 가능한지 알아보았다. 여러 행동 연구 및 신경과학 연구를 살펴 본 결과 순차 처리와 병렬처리는 처리단계에 따라 달라짐을 밝혔다. 초기, 지각적 수준에서는 주의 자원이 유연하게 할당되어 병렬처리가 가능하고 이후 중심단계에서는 순차처리가 일어남을 보였다.

주제어 : 순차처리, 병렬처리