Manuscript accepted in Psychological Science

Evidence of serial processing in visual word recognition

Alex L White*, John Palmer and Geoffrey M. Boynton

*Corresponding author:

Alex White
University of Washington Department of Psychology
119 Guthrie Hall, Box 351525,
Seattle, WA 98195
(206)-543-5161
email: alexlw@uw.edu.
Abstract

To test the limits of parallel processing in vision, we investigated whether people can recognize two words at once. Participants viewed brief, masked pairs of words and were instructed in advance to judge both of the words (dual-task condition), or just one of the words (single-task condition). For judgments of semantic category, the dual-task deficit was so large that it supported all-or-none serial processing: participants could recognize only one word and had to guess about the other. Moreover, participants were more likely to be correct about one word if they were incorrect than correct about the other, which also supports a serial model. In contrast, judgments of text color with identical stimuli were consistent with unlimited-capacity parallel processing. Thus, under these conditions, serial processing is necessary to judge the meaning of words but not their physical features. Understanding the implications of this result for natural reading will require further investigation.
Introduction

Vision begins with parallel processing: the retina and early visual cortex encode many stimulus elements at once, across the visual field. If independent and parallel processing continued all the way through the system, you would be able to perceive multiple objects simultaneously with no cost. At the other extreme, a serial bottleneck would allow only one stimulus to be recognized at a time, causing delays or errors when you must attend to multiple stimuli. Divided attention does often impair task performance (Braun, 1998; Carrasco, 2011), but the impairments are usually too small to be explained by serial processing.

Written words provide an important test of the limits of parallel processing, with clear applications to life in the modern world. Although crowding and poor peripheral acuity require a sequence of eye movements to scan a page, multiple words are still visible within each glance. Indeed, many words in a line of text are at least partially processed while the eyes fixate on the previous word, and some are skipped over by the eyes completely (Rayner, 2009). Within each glance, early visual mechanisms first encode the basic features of many letter shapes in parallel (Grainger, Dufau, & Ziegler, 2016). Eventually, representations of letter combinations lead to full semantic recognition. The present study was motivated by the question: does the visual system’s parallel architecture allow for two words to be processed simultaneously, or is recognition constrained by a serial bottleneck?

The extent of parallel processing in natural reading has been fiercely debated (Murray, Fischer, & Tatler, 2013; Starr & Rayner, 2001). Some researchers argue for strictly serial processing along the line of text (Reichle, Liversedge, Pollatsek, & Rayner, 2009), and others argue for a graded allocation of attention and parallel processing of multiple words (Engbert, Nuthmann, Richter, & Kliegl, 2005; Kennedy, 2000). The parallel/serial debate has proved difficult to resolve with eye movement studies in natural reading.

To determine whether people can recognize two words simultaneously when they are forced to try, we abandoned natural reading conditions and
focused on isolated word recognition during fixation. Our goal was to measure “capacity limits”: constraints on how much information the perceptual system can process per unit time. Previous studies have shown that word recognition is subject to capacity limits by using varieties of search paradigms (Harris, Pashler, & Coburn, 2004; Mullin & Egeth, 1989; Reichle, Vanyukov, Laurent, & Warren, 2008; Scharff, Palmer, & Moore, 2011). None of these visual search studies conclusively determined whether the capacity limit is due to a serial bottleneck (see Discussion).

Rather than visual search, we adopted another classic paradigm: comparing accuracy in dual-task and single-task conditions. In two experiments, we presented participants with time-limited and masked pairs of parafoveal nouns. Participants fixated their gaze in the center and were instructed to detect targets on just one side (single-task condition), or to detect targets on both sides (dual-task condition). We then compared the relative deficit in the dual-task condition to the quantitative predictions of several models of parallel or serial processing.

In separate blocks of trials with identical stimuli and matched difficulty, participants detected either semantic targets or color targets. Semantic targets were nouns that belonged to a particular category, such as ‘professions.’ Color targets were nouns colored slightly reddish, and their semantic meaning was irrelevant. In our “dual-task” conditions, participants always made the same type of judgment for both words (e.g., semantic-semantic or color-color). They never had to make a semantic judgment and a color judgment in the same trial.

Comparing semantic and color judgments allows us to test whether the capacity limit in divided attention depends on which stimulus aspect is task-relevant. Previous studies have demonstrated that detecting changes in low-level features of nonlinguistic stimuli has unlimited capacity (Bonnel, Stein, & Bertucci, 1992; Scharff et al., 2011; White, Runeson, Palmer, Ernst, & Boynton, 2017). However, if word recognition is automatic (Augustinova & Ferrand, 2014; Stroop,
and uses a common resource (Pastukhov, Fischer, & Braun, 2009), then we would predict similar dual-task deficits for both types of judgments.

In the first experiment, participants detected targets embedded in rapid serial visual presentation (RSVP) streams (Figure 1a). RSVP of single words has been studied as a way to present text that doesn’t require saccades (Potter, 1984). On each trial we presented five pairs of unrelated words in RSVP. The presentation rate was adjusted to keep each participant’s single-task semantic performance below ceiling. Fast rates limit the time available to process each pair of words, and reduce the likelihood of a serial shift of attention from one word to the other within one frame. To match color and semantic single-task difficulty levels, we adjusted the saturation of the red targets.

The second experiment tested whether the semantic dual-task deficit in Experiment 1 was due to the RSVP streams overloading memory or other cognitive mechanisms, rather than a limit on the immediate processing of each pair of words. Experiment 2 differed in that each trial contained only one masked pair of words (Figure 1b).

METHODS

Experiment 1

Participants: Ten volunteers (4 female, ages 19-30) with normal or corrected-to-normal visual acuity and normal color vision participated in exchange for fixed monetary payment. All but two (including author AW) were naive as to the research aims. The sample size of 10 was chosen prior to data collection, on the basis of a pilot experiment, to produce standard deviations of the dual-task deficits near 0.01 (in $A_g$ units, see below). This is small relative to the dual-task deficit of 0.08 $A_g$ predicted by the fixed-capacity parallel model described below (assuming equal division of attention).

Each participant gave informed consent in accord with the Declaration of Helsinki and the Institutional Review Board at the University of Washington. All participants learned English as their first language, and all scored above the
norm of 100 ($M \pm SEM: 120 \pm 3$) on the composite Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999).

**Equipment and stimuli:** We presented stimuli on a linearized CRT monitor with 120 Hz refresh rate while the right eye’s gaze position was monitored by an Eyelink eye-tracker (SR Research). The stimuli consisted of: a white background, a small dark fixation cross with dimensions 0.3x0.3 degrees of visual angle ($^\circ$); and dark letter strings in Courier font. The letter strings had 83% Weber contrast. The words were drawn from 12 distinct semantic categories (e.g., “professions”) each with 35 nouns (see Supplementary Materials). The median lexical frequency was 6.4 per million, according to the Clearpond database (Marian, Bartolotti, Chabal, & Shook, 2012). They ranged from 4 to 6 characters in length, subtending 2.7° to 4.1° in width, and 1.1° in height. We also used six-character masks (#@#@#@ and @#@#@@). The words and masks were centered at 2.75° to the left and right of fixation.

All the words were dark gray except for the color targets, which were equiluminant but with higher saturation in the red hue. Using the measured luminance values of the monitor, we incremented the red gun and decreased the green and blue guns to keep the total luminance constant. As described below, we set the magnitude of the red increment to each participant’s detection threshold.

**Trial sequence:** The trial began with the participant fixating centrally for at least 1 s. Then the pre-cue appeared: two 0.35° lines just to the left and right of the fixation cross. In the dual-task condition, both were black. In the single-task condition, one line was green and the other blue. Each participant was assigned to either green or blue and always attended to the side indicted by that color. A target-defining word (the name of the semantic category, or the word “color”) also appeared 1° below fixation.
Figure 1. Example dual-task semantic trials of Experiments 1 (a) and 2 (b). In Experiment 1, the RSVP presentation rate was adjusted for each participant (see text). Not shown in panel (a) are the inter-stimulus intervals (ISIs) between each frame, during which only the fixation cross was presented. In Experiment 2, the duration of the two ISIs between masks and words was adjusted for each participant. The category name (“anatomy”) was actually presented below fixation. In these examples, the post-cue is the blue line that points to the side to be judged, but for half the participants it was the green line. These are dual-task trials, but in single-task trials there was only one post-cue, as the observer had to judge only one side.

The 1 s pre-cue was followed by a 600 ms inter-stimulus interval (ISI). Then the RSVP sequence began with pre-masks that covered the upcoming word locations, followed by a blank ISI containing only the fixation cross. Then 5 pairs of words were presented sequentially, separated by blank ISIs. (The ISIs are not shown in Figure 1). The pre-mask, words, and ISIs all had the same duration, $D_{RSVP}$, which was adjusted to control semantic judgment difficulty (see below). After the last ISI, a post-cue display appeared. This consisted of post-masks and a green and a blue line as in the pre-cue. After 700 ms, a beep
prompted the subject to respond to the side indicated by their assigned color. Responses before the beep were not recorded.

The post-cue remained visible until the participant pressed one of four keys with the hand on the same side as the post-cue, reporting with a 1-4 rating the level of confidence that a target was present: “sure absent”; “guess absent”; “guess present”; or “sure present”. Each key-press was immediately followed a high- or low-pitched feedback tone for correct or incorrect responses, respectively. One dual-task trials, as illustrated in Figure 1, a second post-cue was then presented to prompt a response about the other side.

If the participant’s gaze position moved too far from the fixation mark during the presentation of the stimuli, the trial was immediately terminated (see Supplemental Materials). This occurred on an average of 5.9% of trials (SEM = 1.0%).

Tasks: For semantic judgments, the targets were nouns of a particular semantic category. For color judgments, the targets were words colored slightly red, as defined above. On each trial, on each side, there was an independent 50% chance of semantic target presence, and an independent 50% chance of color target presence. For both target types, therefore, in any given trial there could be no targets, one target, or two targets (one on each side).

The time of each target type within the sequence was uniformly and independently distributed. When there were two targets of the same type, they appeared simultaneously. Only 10% of semantic targets happened to also be color targets. The particular words on each trial were chosen randomly from the entire set with the following constraints:

1. All words on each side were unique.
2. The same word was never presented at both sides simultaneously.
3. No more than one word from the target category could be present on one side.
4. Target words were not allowed to repeat across sequential trials.
In single-task trials, the pre-cue instructed participants to attend to one side and ignore the other. The post-cue always indicated the same side as the pre-cue, requiring just one response. In dual-task trials, participants had to make independent judgments of target presence for both sides. The post-cue first prompted the participant to respond to one side, and then after the key-press and feedback tone, the post-cue switched and the participant responded to the other side. The post-cue order (left or right side first) was randomized across trials. Note again that in dual-task trials, both judgments were of the same type (semantic-semantic or color-color).

Procedure: In the first session, participants read the entire word list and practiced both judgment types. In the second session, they ran staircase blocks to estimate thresholds. A staircase for single-task semantic judgments established the threshold word duration $D_{RSVP}$. Then, with the stimulus timing fixed, a staircase for single-task color judgments established the threshold color increment magnitude $I$. The average $D_{RSVP}$ was $83 \pm 8.5$ ms ($M \pm SEM$), corresponding to a presentation rate of $6.7 \pm 0.7$ Hz. The average $I$ (expressed as a proportion of the maximum possible equiluminant red saturation) was $0.09 \pm 0.01$.

Main experimental trials were run in blocks of 16. The target type (semantic or color) was constant within each block, as was the semantic target’s category and the pre-cue condition (dual-task; single-task right; single-task left). Each participant completed 960 trials (60 blocks) of each judgment type.

Analysis: As a bias-free measure of accuracy, we computed the area under the Receiver Operating Characteristic (ROC) curve, known as Green’s Area or $A_g$ (Pollack & Hsieh, 1969). Like proportion correct, $A_g$ is a proportion that ranges from 0.5 (guessing) to 1.0 (perfect). See Supplementary Materials for details.
Experiment 2

Participants: Ten volunteers participated (3 female, ages 19-31). Eight had also participated in Experiment 1. One additional participant chose to discontinue the study after one session. The mean TOWRE score was 119 ± 2.

Trial sequence & Procedure: The methods of Experiment 2 (Figure 1b) were identical to Experiment 1 except as noted here. After the pre-cue, a pre-mask display appeared for 42 ms. It consisted of two strings of 6 randomly chosen consonants. Then only the fixation mark was presented for an ISI, the duration of which (D_{ISI}) was set to each subject’s threshold for single-task semantic judgments. Then a pair of words appeared for 42 ms, followed by a second ISI of the same duration of the first. The following post-mask was similar to the pre-mask, 42 ms in duration, but with different random consonants. After the post-mask, the trial sequence finished in the same way as in Experiment 1. Trials were run in blocks of 20.

There was an independent 50% chance that the word on each side was drawn from the target semantic category. Otherwise, it was drawn randomly from one of the other 11 categories. No word was allowed to appear on two consecutive trials, or to appear on both sides in the same trial. There was also an independent 50% chance that each word was a color target.

The average D_{ISI} was 71 ± 6.2 ms, and the average color target saturation increment I was 0.14 ± 0.01. An average of 5.1% of trials were aborted due to fixation breaks (SEM = 1.7%).

RESULTS

Dual-task deficits:

Semantic accuracy was significantly impaired in the dual-task conditions of both experiments (Table 1). The mean dual-task deficit (single-task minus dual-task A_g) was 0.12 ± 0.01 in Experiment 1 (comparison to 0: t(9)=17.17, p<0.001; 95% bootstrapped CI = [0.11 0.13]), and 0.14 ± 0.01 in Experiment 2 (t(9)=10.14, p<0.001; CI = [0.12 0.17]). Color judgments, in contrast, suffered minimal deficits:
0.02 ± 0.01 in Experiment 1 (t(9)=1.55, p=0.15; CI = [0.0 0.04]) and 0.04 ± 0.01 in Experiment 2 (t(9)=4.63, p=0.001; CI = [0.02 0.06]). Dual-task deficits were significantly greater for semantic than color judgments, according to ANOVA interactions between judgment type and pre-cue condition (Experiment 1: F(1,9) = 54.48, p<0.001; Experiment 2: F(1,9) = 30.73, p<0.001). The same analysis of d' (which, unlike A_g, is unbounded) supported the same conclusion (interactions: F(1,9) = 28.06, p<0.001; F(1,9)=40.21, p<0.001).

Accuracy was consistently higher for targets on the right than left side of fixation, especially for semantic judgments consistent with previous reports (e.g., Boles, 1983; Mishkin & Forgays, 1952). Details are in the Supplementary Materials.

Table 1: mean accuracy levels (A_g) in single- and dual-task conditions. In all cells, SEM = 0.01.

<table>
<thead>
<tr>
<th></th>
<th>Experiment 1</th>
<th></th>
<th>Experiment 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>semantic</td>
<td>color</td>
<td>semantic</td>
<td>color</td>
</tr>
<tr>
<td>single-task A_g</td>
<td>0.83</td>
<td>0.82</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>dual-task A_g</td>
<td>0.71</td>
<td>0.80</td>
<td>0.70</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Attention Operating Characteristics:

The “Attention Operating Characteristic” (AOC) allows the comparison of dual-task deficits to specific model predictions (Sperling & Melchner, 1978). 

**Figure 2** plots accuracy for words on the left side of fixation against accuracy for words on the right. The single-task conditions are pinned to their respective axes. The accuracy levels in the dual-task condition form a single point (open symbol) in that 2D space. We then compare that point to the predictions of three specific models of capacity limits (Bonnel & Prinzmetal, 1998; Scharff et al., 2011; Shaw, 1980; Sperling & Melchner, 1978). Below they are described in brief, with more details in the Supplementary Materials.

1. *Unlimited-capacity parallel processing*: two stimuli can be fully processed simultaneously just as well as one (i.e., no dual-task deficit). In the AOC, this
model predicts that the dual-task data point falls at the intersection of the dashed lines.

2. **Fixed-capacity parallel processing:** The perceptual system extracts a fixed amount of information from the whole display per unit time. Therefore, processing resources must be shared between both stimuli in the dual-task condition, which lowers sensitivity. As the proportion of resources given to the right stimulus increases from 0 to 1, this model traces out the black curve in the AOC plot. See the **Supplementary Materials** for the calculation of this curve.

3. **All-or-none serial processing:** only one stimulus can be processed per trial, with equal sensitivity as in the single-task condition. The participant does not have time to even start processing the other stimulus and therefore must guess when asked about it. As the proportion $v$ of trials in which the right side is processed increases from 0 to 1, this model traces out the diagonal black line in the AOC plot.

The serial model can be generalized to account for less severe deficits by assuming that on some fraction $b$ of dual-task trials, both sides are fully processed. The resulting accuracy is a mixture of trials in which only one stimulus is processed and trials in which both stimuli are processed. In conditions with dual-task deficits larger than predicted by either parallel model, we solved for the serial model parameters $b$ and $v$ that best fit each participant’s data.
Figure 2. Attention Operating Characteristics (AOCs) for semantic judgments (left column) and color judgments (right column) in Experiments 1 (top row) and 2 (bottom row). Solid points pinned to the axes are single-task accuracy levels ($A_g$); open points are dual-task. Error bars show ± 1 SEM.

**Semantic judgments, Experiment 1:** As can be seen in Figure 2a, dual-task accuracy (open symbol) fell below the fixed-capacity parallel model’s curve, and near but above the all-or-none serial model’s line. The mean distance to the nearest point on the fixed-capacity parallel curve was 0.05 ± 0.01 (significantly greater than 0: $t(9) = 4.53, p=0.001; CI = [0.03 0.07]$). The mean distance to the all-or-none serial line was 0.04 ± 0.01 ($t(9) = 4.87, p<0.001; CI = [0.03 0.06]$). Fitting the more general serial model, the mean $b$ parameter was 0.20 ± 0.04 (CI = [0.14 0.28]), meaning that on an average of 80% of the trials only one stimulus was processed. The mean $\nu$ was 0.71 ± 0.05 (CI = [0.62 0.82]), meaning that
there was a significant bias to process the right stimulus when only one could be processed. This bias might reflect an attentional strategy to process the easier side, given that only one could be processed successfully.

**Semantic judgments, Experiment 2:** Mean dual-task accuracy fell on top of the all-or-none serial model’s prediction (Figure 2c). The average distance from that line was $0.005 \pm 0.012$, not significantly above 0 ($t(9) = 0.41, p=0.69; CI = [0.02 0.03]$). The average distance to the closest point on the fixed-capacity parallel curve was $0.08 \pm 0.02$ ($t(9) = 4.52, p=0.001; CI = [0.05 0.12]$). The mean value of $b$ was $0.06 \pm 0.03$ ($CI = [0.01 0.14]$), meaning that on 94% of trials only one stimulus was processed. There was again a strong bias to process the right stimulus: mean $v = 0.79 \pm 0.04$ ($CI = [0.71 0.86]$).

**Color judgments, Experiment 1:** The mean dual-task accuracy point fell well above the fixed-capacity parallel curve, near the intersection that marks the prediction of the unlimited-capacity parallel model (Figure 2b). The mean distance to the nearest point on the fixed-capacity curve was $0.09 \pm 0.01$, significantly above 0 ($t(9) = 7.07, p<0.001; CI = [0.06 0.11]$).

**Color judgments, Experiment 2:** As in Experiment 1, dual-task accuracy fell near the unlimited-capacity parallel model’s prediction (Figure 2d). The mean distance from the fixed-capacity parallel curve was $0.06 \pm 0.01$ ($t(9) = 6.25, p<0.001; CI=[0.04 0.07]$).

**Effects of accuracy on the other side**

The all-or-none serial model assumes that only one side can be processed per trial and no information is acquired about the other. If we also assume that the focus of attention switches across trials between the left and right sides, the model predicts a negative correlation between the accuracies of dual-task responses (Bonnel & Prinzmetal, 1998; Ernst, Palmer, & Boynton, 2012; Lee, Koch, & Braun, 1999; Sperling & Melchner, 1978). In other words, the participant is more likely to be correct about one side when s/he is incorrect about the other side.
There are several ways to test this prediction, including computing correlation coefficients between the accuracies of the left- and right-side responses. For semantic judgments, the correlation coefficients were negative except when neither side contained a target (see Supplementary Materials). This complex pattern seemed related to changes in decision criterion or bias as a function of the other side’s task. We therefore need a bias-free measure of accuracy, such as area under the ROC curve ($A_g$). So, as another direct test of the serial switching model’s prediction (Braun & Julesz, 1998; Lee et al., 1999), we coded all dual-task responses by whether the response to the other side on the same trial was correct or incorrect, and then computed $A_g$ for both sets of trials (Figure 3).

![Figure 3](image)

**Figure 3.** Dual-task accuracy ($A_g$) when the other side’s response was incorrect plotted against accuracy when the other side’s response was correct. (a) Experiment 1; (b) Experiment 2. Each point is one participant’s data for one judgment type (semantic judgments = blue circles; color judgments = red diamonds). Points below the unity line indicate that accuracy was better when the other side’s response was incorrect.

For semantic judgments in Experiment 1, accuracy was on average $0.06 \pm 0.02$ $A_g$ units lower when the other side’s response was correct than incorrect ($t(9)=3.31, p=0.009; CI = [0.03 0.11]$). In Experiment 2, this difference was $0.11 \pm$
This pattern supports the serial switching model.

For color judgments, the effect was exactly opposite: In Experiment 1, accuracy was on average 0.07 ± 0.02 Ag units higher when the other side’s response was correct than incorrect ($t(9)=3.60, p=0.006; \text{CI} = [0.03 0.10]$). The same pattern occurred in Experiment 2: the mean difference was 0.04 ± 0.02 ($t(9)=2.99, p=0.015; \text{CI} = [0.02 0.08]$). This positive effect of the other side’s accuracy is consistent with fluctuations in overall effort or arousal that could cause positive correlations in sensitivity between the two sides (Bonnel & Prinzmetal, 1998).

In both experiments, the effect of other side’s accuracy significantly differed between semantic and color judgments (ANOVA interactions in Experiment 1: $F(1,9)=30.63, p<0.001$; Experiment 2: $F(1,9)=57.02, p<0.001$).

**DISCUSSION**

We measured dual-task deficits for color and semantic judgments of written words. For color judgments, there was minimal deficit, consistent with unlimited-capacity parallel processing. For semantic judgments, there was a large dual-task deficit that was inconsistent with unlimited-capacity or even fixed-capacity parallel processing. Instead, it supported an all-or-none serial model: a bottleneck in the recognition process that allows only one word to be categorized per trial. Given limited processing time, the serial model predicts that participants process only the left stimulus on some trials and only the right on others. Fulfilling that prediction, semantic accuracy for each side was relatively impaired when the other side was judged correctly. (Note that our model fits suggested that participants could process both words on a minority of trials – 20% and 6% in Experiments 1 and 2, respectively). The opposite was true for color judgments. Based on these contrasting results, we argue that serial processing is necessary to judge the meaning of words but not their physical features.
The consistently large semantic dual-task deficit is unlikely to be explained by a memory limit, for several reasons. First, participants in Experiment 2 had to remember only two words for the brief interval between the stimuli and the post-cue. Two words is within the limits of verbal working memory (Chen & Cowan, 2009). Second, dual-task color trials also required two reports with the same timing and suffered hardly any deficit. Finally, accuracy of the second responses on dual-task trials was not worse than the first responses, suggesting that both stimuli could be remembered for the whole trial. We therefore favor the hypothesis that the serial bottleneck lies in perceptual and/or linguistic analysis.

The color and semantic data fell at the opposite extremes despite the fact that the stimuli and single-task difficulty levels were the same. Therefore, the extent of parallel processing is not fixed by the stimulus but depends on the demands of the task. This finding also casts doubt on the view that semantic processing is automatic (Augustinova & Ferrand, 2014; Stroop, 1935). If both words were automatically processed in single-task semantic trials, accuracy would have been no better than in dual-task trials. Moreover, the lack of dual-task deficit for color judgments shows that participants can prevent semantic processing of the two attended words from overwhelming a central pool of limited resources. This supports the hypothesis that semantic processing requires attention and is under top-down control (e.g., Robidoux & Besner, 2015). Consistent with this view, a recent study demonstrated that when pairs of Chinese words are superimposed in different colors, readers are able to attend to one word at a time and avoid interference from the other (Liu & Reichle, in press.)

Although the color data were consistent with parallel processing, our design was not optimized to rule out a rapid switching of attention that sequentially processed both colors, because the stimulus timing was set to control semantic difficulty only. The more predictable color target features could also account for the difference in results. Nonetheless, we show that different tasks of equal difficulty performed on the same stimuli can have very different capacity limits.
It is likely that the large semantic dual-task deficit depended on our use of displays that limited the amount of time available to process each word pair. By the time one word was recognized, information about the other was lost or replaced by subsequent stimuli. In contrast, another study from our group using a semantic search task without post-masks found results consistent with the fixed-capacity parallel model (Scharff et al., 2011). The lack of masking may have allowed multiple words to be processed sequentially. We propose that recognition is always serial, but multiple words can be processed in a single presentation if the stimuli or their memory traces last long enough.

Prior studies have demonstrated semantic capacity limits with evidence that response times increase with set size during search for particular words (Harris et al., 2004; Reichle et al., 2008). Reichle et al. (2008) further argued that visual features are detected in parallel but lexical processing is serial. However, search slopes alone are insufficient to distinguish between serial and parallel models (Carrasco & Yeshurun, 1998; Townsend, 1990).

The “redundant target” paradigm has also been used to investigate capacity limits in word recognition. Assuming that search is self-terminating, unlimited-capacity parallel models predict faster responses when the display consists of two targets than a single target. The first such study found no redundancy gains and concluded that capacity is limited, potentially (but not necessarily) due to a serial process (Mullin & Egeth, 1989). However, later studies reported contradictory results (Shepherdson & Miller, 2014, 2016).

More generally, the signatures of serial processing have rarely been observed in dual visual tasks. The studies that have observed them involved task conditions more complex than ours (Bonnel & Prinzmetal, 1998; Braun & Julesz, 1998; Lee et al., 1999; Sperling & Melchner, 1978). In all of these cases, the ‘attentional set’ (i.e., which stimulus features the subject looks for and discriminates) was different for the two concurrent tasks. Our study is therefore unique: the left- and right-side tasks were identical.
Reading researchers have long debated whether words in a line of text are necessarily processed one at a time (Murray et al., 2013; Reichle et al., 2009). Much of the debate has focused on eye movement studies that demonstrate that readers begin processing word \( n+1 \) before the eyes leave word \( n \). There is even evidence that the \textit{meaning} of word \( n+1 \) is acquired before it is fixated (Hohenstein & Kliegl, 2014; Schotter, 2013), which is consistent with parallel models. However, proponents of serial models argue that they can also account for such semantic effects (Schotter, Reichle, & Rayner, 2014). Our study avoids that impasse and demonstrates a boundary condition in which parallel recognition of two words is not possible.

The conditions of our study differed from natural reading, so we cannot say with certainty how our findings apply to the processing of a whole page of text. By demonstrating one condition in which semantic processing of two words is strictly serial, we do not rule out the hypothesis that parallel processing is at least sometimes possible during natural reading. For instance, parallel processing might be possible when pairs of words are related to each other, or when one word is fixated and the other is to the right. Reading is a remarkable skill, and future research must explore a wider range of tasks and stimulus configurations to map out the cognitive functions that support it.

**Acknowledgments:**

This work was supported by NEI grants T32 EY007031 to the University of Washington, R01 EY12925 to authors GMB and JP, and F32 EY026785 to author AW. Thanks to Kadie Anne Brown, Micah Ketola and Evan McAuley for help with data collection.

**Author contributions**

ALW, JP and GMB designed the study; ALW collected and analyzed the data; ALW wrote the manuscript, and ALW, JP and GMP edited it. All authors approved the final version of the manuscript.
REFERENCES


of Vision, 12, 1–17.


Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-


SUPPLEMENTARY MATERIALS (SOM-R)

For
White, Palmer & Boynton: Evidence of serial processing in visual word recognition

SUPPLEMENTARY METHODS

Screen and eye-tracking details

In Experiment 1, the screen resolution was 832 x 624 pixels, with maximum luminance 104 cd/m². In Experiment 2, the screen resolution was 1024 x 640 pixels, with maximum luminance 90 cd/m². The luminance of the words was 21 cd/m² in Experiment 1 and 18.9 cd/m² in Experiment 2. Stimuli were created with Matlab software (MathWorks) and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

In both experiments we monitored the right eye’s gaze position with Eyelink eye-trackers (a head-mounted Eyelink 2 and an Eyelink 1000; SR Research). Fixation was established during the ITI at the start of each trial. The trial only advanced if the estimated gaze position was within 1.5 degree of visual angle (°) horizontally and 4° vertically of the fixation cross for at least 200 ms. We allowed more vertical tolerance to accommodate drifts due to pupil size changes or slippage of the eye-tracker. The gaze position averaged over the next 10 samples was taken as the current trial’s fixation position. A fixation break was then defined as a deviation of gaze position more than 1° horizontally or 2° vertically from the fixation position established at the trial’s start. If a fixation break occurred during the RSVP sequences in Experiment 1, or between the pre-cue offset and post-mask offset in Experiment 2, the trial was immediately terminated. The participant had to press a button to continue the next trial. Terminated trials were repeated at the end of the block, unless fewer than 3 trials remained.

Procedure
In staircase blocks, we used the “single-interval adjustment matrix” adaptive procedure (Kaernbach, 1990) to find thresholds midway between chance and perfect detection performance. The attended side (left, right) alternated across staircase blocks.

In each hour-long session after the staircase session, participants completed 12 semantic blocks (one for each category) and 12 color blocks, alternating order across sessions. The difficulty levels $D$ and $I$ were adjusted as necessary between sessions to keep performance near 80% correct in the single-task conditions. We discarded and re-ran any set of 12 blocks if both single- and dual-task proportion correct were below 0.7 or above 0.9. This occurred if stimulus levels $D$ or $I$ missed the participant’s true threshold. As a result, a total of 96 and 84 blocks were re-run in Experiments 1 and 2, respectively. Each participant completed 960 usable trials of each judgment type.

**Analysis**

On each trial the participant reported his/her level of confidence that a target was present on a 1-to-4 rating scale. As a bias-free measure of accuracy, we computed the area under the Receiver Operating Characteristic (ROC) curve, $A_g$ (Pollack & Hsieh, 1969). The ROC plots hit rates (HR) as a function of false alarm rates (FR). To compute these rates from the participants’ response ratings, we varied an index $i$ from 0 to 4. At each index level we coded responses greater than $i$ as “yes” responses. For each value of $i$, $HR(i)$ is the proportion of “yes” responses on target-present trials and $FR(i)$ is the proportion of “yes” responses on target-absent trials. For instance, when $i=3$, only response ratings of 4 (highest confidence) on target-present trials are considered hits, and only response ratings of 4 on target-absent trials are considered false alarms. The five pairs of $HR(i)$ and $FR(i)$ trace out a curve, the area under which ($A_g$) is a measure of accuracy.

Throughout the text we report bootstrapped 95% confidence intervals (CIs) for average measurements. To compute these, we generate a distribution
of resampled means. Each of those is the mean of 10 values sampled with
replacement from the original set of 10 participants. The CI is the range from the
2.5\textsuperscript{th} to 97.5\textsuperscript{th} percentile of the distribution of resampled means.

Models of capacity limits

In the Attention Operating Characteristic plots shown in Figure 2, we compare
our dual-task accuracy data to the quantitative predictions of three models of
capacity limits. The following definitions are for "standard" versions of these
models similar to those defined for accuracy measures of visual search tasks
(Scharff, Palmer, & Moore, 2011). All models are based on signal detection
theory: we assume that the perceptual system analyzes stimuli on each side \(i\) by
computing an estimate of the evidence that a target is present, which
 corresponds to a random variable \(E_i\). In all models, we assume that across trials
of one condition, the \(E_i\) values for side \(i\) are independent and identically
distributed Gaussian variables, and within each trial \(E_1\) and \(E_2\) are independent of
each other.

Sensitivity (\(d'\)) depends on the mean difference in \(E\) between target-present
and target-absent trials, relative to the across-trial variability in \(E\). To make a
judgment about each stimulus, \(E\) is compared against three criteria \(c1, c2, c3\) to
determine which of the four response keys to press.

We first label the measured single-task accuracy levels for the left and right
stimuli \(A_{L1}\) and \(A_{R1}\), respectively. The models then use these single-task accuracy
levels to predict the dual-task levels \(A_{L2}\) and \(A_{R2}\).

4. Unlimited-capacity parallel processing model: This model assumes that the
distribution of \(E_i\) is independent of the condition (single-task vs. dual-task).
The model therefore predicts no dual-task deficit:

\[A_{L2} = A_{L1},\] and
5. **Fixed-capacity parallel processing model**: This model assumes that the perceptual system extracts a constant amount of information from the display regardless of the condition (single-task vs. dual-task). If that information is equally distributed among the stimuli in the dual-task condition, only half as much information is available about each stimulus compared to the single-task condition. One way to conceptualize fixed capacity is to assume that computing the estimate $E_i$ of target presence depends on gathering sensory ‘samples’ from the stimulus (Shaw, 1980). All attended stimuli must share a fixed number $S$ of samples that can be gathered from the whole display per unit time. The variability of $E_i$ is inversely proportional to the number of samples assigned to it, which means that reducing the number of samples for one stimulus decreases sensitivity. As the proportion $q$ of samples given to the right stimulus increases from 0 to 1, this model’s prediction traces out the black curve that connects the two single-task data points in the AOC plot. This curve is computed as follows. We first calculate the value of $d'$ for the left and right single-task conditions:

$$d'_{R1} = \sqrt{2} z(A_{R1}),$$
$$d'_{L1} = \sqrt{2} z(A_{L1})$$

where $z$ is the inverse of the normal cumulative distribution function.

We then assume that in the dual-task condition, the right stimulus receives $qS$ samples, and the left stimulus receives the remaining $(1-q)S$ samples, where $0 < q < 1$. From signal detection theory, receiving a portion $q$ of samples changes $d'$ for that stimulus by a factor $\sqrt{q}$. Therefore, $d'$ for each side in the dual-task condition is:

$$d'_{R2} = \sqrt{q} d'_{R1},$$
$$d'_{L2} = \sqrt{1-q} d'_{L1}$$

We then convert these d-prime measures into $A_g$ accuracy levels:
\[ A_{R2} = ncdf\left(\frac{d'_{R2}}{\sqrt{2}}\right), \text{ and} \]
\[ A_{L2} = ncdf\left(\frac{d'_{L2}}{\sqrt{2}}\right) \]

where \( ncdf \) is the normal cumulative distribution function.

The parallel model can be generalized to predict less severe deficits, by assuming that the total number of samples available in the dual-task condition is more than the number \( S \) in the single-task condition. That is, the total number of samples shared between the two locations is \( aS \), where \( 1 \leq a \leq 2 \). Increasing \( a \) pushes the predicted curve into the upper right corner, eventually meeting the prediction of the unlimited capacity model when \( a = 2 \).

6. **All-or-none serial processing model**: This model assumes the same stimulus representations and decision rule as the parallel models. Like in the unlimited-capacity parallel model, the distributions of \( E \) are identical in single- and dual-task conditions. The difference is that this model assumes that the participant only processes one stimulus and does not have time to even start processing the second, and therefore must guess when asked about it. As the proportion of trials \( v \) in which the right side is processed increases from 0 to 1, this model’s prediction traces out the diagonal black line in the AOC plot. Accuracy values along that line, which are in units of \( A_g \) but may also be viewed as the proportion of correct responses, are therefore a mixture of probabilities. Dual-task accuracy for the right stimulus therefore is:

\[ A_{R2} = A_{R1} v + 0.5(1-v). \]

The second term in that equation reflects the fact that the participant must guess (with probability correct 0.5) on the \( (1-v) \) proportion of trials in which the right stimulus is not processed at all. The right and left sides trade off linearly, so dual-task accuracy for the left stimulus is:

\[ A_{L2} = A_{L1}(1-v) + 0.5v. \]

The serial model can be generalized to account for less severe deficits assuming that on some fraction \( b \) of dual-task trials, both sides are fully
processed (with the same sensitivity as in the single-task conditions). Therefore, the resulting dual-task accuracy is a mixture of trials in which only one stimulus is processed and no information is acquired about the other, and in which both stimuli are fully processed.

\[ A_{R2} = bA_{R1} + (1-b)*(A_{R1}v + 0.5(1-v)) \]
\[ A_{L2} = bA_{L1} + (1-b)*(A_{L1}(1-v) + 0.5v) \]

In this generalized model, \( v \) can be interpreted as the proportion of trials in which the right stimulus is processed ‘first’. On only \( b \) fraction of trials does the ‘second’ stimulus get processed at all. We were able to solve these equations for the values of \( b \) and \( v \), given each participant’s data (single-task accuracy values \( A_{L1} \) and \( A_{R1} \), and dual-task values \( A_{L2} \) and \( A_{R2} \)).

**SUPPLEMENTARY RESULTS**

**Effect of response order in dual-task trials**

On dual-task trials, the post-cues prompted participants to report judgments of both sides, in random order. For both judgment types in Experiment 1 and color judgments in Experiment 2, the mean accuracy of second responses was 0.01 ± 0.01 higher than the first, but that difference was not significant (all \( p>0.3 \)). For semantic judgments in Experiment 2, second responses were on average 0.03 ± 0.01 \( A_g \) units more accurate (\( t(9)=4.04, p=0.003 \)). Because these differences were small, all other analyses collapse across first and second dual-task responses. It is likely that these order effects were small because the order was unpredictable (Ernst, Palmer, & Boynton, 2012). For a contrasting result in a dual task, see Egeth & Pachella, 1969.

**Effect of hemifield**

In Experiment 1, overall semantic accuracy was on average 0.16 ± 0.02 \( A_g \) units higher for targets on the right than left side (\( t(9)=9.50, p<0.001; CI = [0.13 \ 0.19] \)). The magnitude of that difference did not significantly vary across single-task and dual-task conditions (ANOVA interaction: \( F(1,9)=3.56, p=0.092 \)).
Experiment 2, the mean hemifield difference was $0.23 \pm 0.02$ ($t(9)=12.5$, $p<0.001$; CI = [0.20 0.27]). That difference was greater in the dual-task ($0.28 +/- 0.02$) than in the single-task condition ($0.15 +/- 0.02$; ANOVA interaction: $F(1,9)=60.47$, $p<0.001$). However, that interaction was not significant in units of $d'$, which is more appropriate as it is unbounded (Experiment 1: $F(1,9)<1$; Experiment 2: $F(1,9)=3.42$, $p=0.097$). Therefore, even when instructed to attend to only one side, participants are less accurate at categorizing words in the left hemifield are than in the right. Therefore, if an attentional bias explains the hemifield asymmetry in general, it must be difficult to voluntarily overcome.

The effect of hemifield almost disappeared for color judgments: within the single-task conditions, there was a significant interaction between judgment type and side (Experiment 1: $F(1,9)=29.37$, $p<0.001$; Experiment 2: $F(1,9) = 20.11$, $p=0.0015$). The same analyses of $d'$ supported the same conclusions. For color judgments in Experiment 1, accuracy was on average $0.03 \pm 0.01$ $A_g$ units higher on the right than the left ($t(9)=2.74$, $p=0.023$; CI = [0.01 0.05]). In Experiment 2, that mean difference was $0.04 \pm 0.02$ $A_g$ units ($t(9)=2.17$, $p=0.058$; CI = [0.0 0.07]). In neither experiment did the color judgment side difference interact with pre-cue condition (single vs. dual; $F<1$).

**Dual-task accuracy correlations**

To test the prediction of the serial model that assumes across-trial switching of attention between the left and right sides, we computed linear correlation coefficients ($\rho$) between the accuracies of the two responses in dual-task trials of each judgment type.

*Semantic judgments:* across all trials of Experiment 1, the mean $\rho$ was negative but not significantly below zero: $-0.02 \pm 0.01$ ($t(9) = -1.23$, $p=0.25$; CI = [-0.04 0.01]). We reasoned, however, that because this task requires yes/no target detection and most participants had a conservative bias, the serial processing bottleneck might affect only the detection of targets. We therefore split our data into sets of trials depending on whether any target was present or absent. When
no targets were present on either side, mean $\rho$ was positive: 0.22 ± 0.03 ($t(9) = 6.13, p<0.001; CI = [0.16 0.29]$). When at least 1 target was present, the mean $\rho$ was negative: -0.11 ± 0.02 ($t(9) = 6.17, p<0.001; CI = [-0.15 -0.08]$).

In Experiment 2, across all semantic trials the mean $\rho$ was not again significantly different from zero: -0.02 ± 0.02 ($t(9) = 1.28, p=0.234; CI = [-0.05 0.01]$). When no targets were present on either side, the mean $\rho$ was positive but not significantly so: 0.04 ± 0.05 ($t(9) = 0.81, p=0.44; CI = [-0.06 0.14]$). On trials with 1 or 2 targets, mean $\rho$ was negative: -0.06 ± 0.03 ($t(9) = 2.07, p=0.068; CI = [-0.11 -0.01]$).

Although the semantic accuracy correlations were negative when at least one target was present, the basic prediction of the serial switching model was not met across all trials. This may be because the overall accuracy correlation collapses over hits and correct rejections, and over misses and false alarms. It is therefore corruptible by shifts in criterion or response bias. To investigate this possibility, we computed a measure of bias:

$B = -0.5*(z(HR) + z(FR))$

where $z$ inverse of the unit normal cumulative distribution function, $HR$ is the hit rate, and $FR$ is the false alarm rate. In the framework of signal detection theory, $B$ is the distance of the decision criterion from the neutral point midway between the distributions of sensory evidence on target-absent and target-present trials, in units of the standard deviation. A larger $B$ means the participant is more conservative, needing more evidence to say yes.

Indeed, $B$ depended on the accuracy of the other side’s response. For semantic judgments in Experiment 1, participants were more conservative (higher $B$) when the other side’s response was correct than incorrect (mean difference: 0.25 ± 0.05, $t(9)=5.11, p<0.001; CI = [0.16 0.34]$). That difference was larger when the other side had no target (0.57 ± 0.11, $t(9)=4.99, p<0.001; CI = [0.38 0.82]$) than when it did (0.04 ± 0.15, $t(9)=0.28, p=0.78; CI = [-0.25 0.30]$). The interaction ($F(1,9)= 4.92, p=0.054$) between other side accuracy and other side target presence corresponds to the difference in accuracy correlations
between target-present and absent trials, with a significantly positive correlation when neither side had a target.

For semantic judgments in Experiment 2, participants were again more conservative when the other side’s response was correct (mean difference in $B = 0.25 \pm 0.05$, $t(9)=4.94$, $p<0.001$; CI = [0.14 0.34]), but that effect did not strongly differ between target-present and absent trials ($F(1,9)<1$). Correspondingly, the accuracy correlations in Experiment 2 differed less between target present and absent trials.

In summary, the analysis of accuracy correlations is complicated by changes in decision bias that depend on the other side’s response and the other side’s stimuli. We therefore rely on the analysis of bias-free accuracy ($A_g$) presented in the main text.

Color judgments showed a very different pattern of accuracy correlations. In Experiment 1, the correlation was overall positive (mean $\rho=0.07 \pm 0.02$; $t(9) = 3.97$, $p=0.003$; CI = [0.04 0.10]), regardless of whether there were (0.05 +/- 0.01) or were not (0.14 +/- 0.08) any color targets present. In Experiment 2, the color accuracy correlation was again overall positive (mean $\rho=0.05 \pm 0.02$, $t(9) = 3.23$, $p=0.010$; CI = [0.02 0.08]) regardless of whether there were (0.04 +/- 0.02) or were not (0.09 +/- 0.04) any color targets present. These positive accuracy correlations mirror the positive effect of the other side’s accuracy on $A_g$, as shown in the main text, and could be explained by fluctuations in overall effort or arousal across trials. The bias ($B$) of color judgments was not significantly affected by the other side’s accuracy (mean effects in Experiment 1: $0.07 \pm 0.05$, $t(9)=1.52$, $p=0.164$, CI = [-0.01 0.16]; in Experiment 2: $0.06 \pm 0.04$, $t(9)=1.66$, $p=0.131$, CI = [-0.01 0.14]).

**Congruency effects**

A congruency effect is the influence of one stimulus on the participant’s response to another stimulus. The classic “flanker effect” (Eriksen & Eriksen, 1974) is an example of a congruency effect: participants are instructed to
discriminate a target stimulus that is flanked by irrelevant stimuli that may be congruent (correspond to the same response) or incongruent (correspond to the opposite response). Responses are typically impaired on incongruent trials, which is a sign that the flankers were not completely filtered out.

We compared accuracy on 'congruent' trials, when both sides have a target or both don’t have a target, with accuracy on 'incongruent' trials, when only one side has a target. For semantic judgments, accuracy was higher on congruent than incongruent trials, by an average of 0.05 ± 0.01 A_g units in Experiment 1 (t(9)=4.23, p=0.002; CI = [0.03 0.08]), and 0.03 +/- 0.01 in Experiment 2 (t(9)=4.62, p=0.001; CI = [0.02 0.04]). Therefore, in addition to their inability to recognize both words simultaneously, participants were not perfectly able to maintain separate representations of the two stimuli and select the correct one to respond to.

Congruency effects were absent or reversed for color judgments: accuracy tended to be higher on incongruent trials. In Experiment 1, that mean difference was 0.03 ± 0.02, but not significant (t(9)=1.18, p=0.27; CI = [-0.02 0.07]). It was larger in Experiment 2: 0.06 ± 0.01 (t(9)=5.49, p<0.001; CI = [0.04 0.08]). This may be because on incongruent trials of Experiment 2, participants directly compared saturation levels across space to better detect changes from the baseline gray. In contrast, in Experiment 1, participants were perhaps more likely to compare colors of successive RSVP stimuli across time.
SUPPLEMENTARY REFERENCES


