Detecting meaning in RSVP at 13 ms per picture

Detecting meaning in RSVP at 13 ms per picture

Mary C. Potter, Brad Wyble, Carl Erick Hagmann and Emily S. McCourt

Massachusetts Institute of Technology

Author Note

Mary C. Potter, Carl Erick Hagmann, and Emily S. McCourt, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

Brad Wyble is now at the Department of Psychology, Pennsylvania State University, University Park, PA 16802.

This research was supported by a National Institutes of Health Grant MH47432.

Correspondence concerning this article should be addressed to Mary C. Potter, Department of Brain and Cognitive Sciences, 46-4125, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139. E-mail: mpotter@mit.edu

Detecting meaning in RSVP at 13 ms per picture

Abstract

The visual system is exquisitely adapted to the task of extracting conceptual information from visual input with every new eye fixation, three or four times a second. Here we assess the minimum viewing time needed for visual comprehension, using rapid serial visual presentation (RSVP) of a series of 6 or 12 pictures presented between 13 and 80 ms per picture with no inter-stimulus interval. Participants detected a picture specified by a name (e.g., *smiling couple*) that was given just before or immediately after the sequence. Detection improved with increasing duration and was better when the name was presented before the sequence, but performance was significantly above chance at all durations, whether the target was named before or only after the sequence. The results are consistent with feedforward models in which an initial wave of neural activity through the ventral stream is sufficient to allow identification of a complex visual stimulus in a single forward pass. Although other explanations are discussed, the results suggest that neither reentrant processing from higher to lower levels nor advance information about the stimulus is necessary for the conscious detection of rapidly presented, complex visual information.

190 words

Keywords: picture perception, feedforward processing, attentional set, conscious perception, conceptual processing

.

Detecting meaning in RSVP at 13 ms per picture

Our eyes move to take in new information three or four times a second, and our understanding of visual input seems to keep pace with this information flow. Eye fixation durations may be longer than the time required to perceive a scene, however, because they include time to encode the scene into memory and to plan and initiate the next saccade. Indeed, a picture as brief as 20 ms is easy to see if it is followed by a blank visual field (e.g., Thorpe, Fize, & Marlot, 1996). However, presenting another patterned stimulus after the target as a mask interferes with processing, particularly if the mask is another meaningful picture (Potter, 1976, Intraub, 1984, Loftus, Hanna, & Lester, 1988, Loschky, Hansen, Sethi, & Pydimarri, 2010). With rapid serial visual presentation (RSVP) of colored photographs of diverse scenes, each picture masks the preceding one, so only the last picture is not masked. Nonetheless, viewers can detect a picture presented for 125 ms in an RSVP sequence when they have only been given a name for the target, such as *picnic* or *harbor with boats* (Intraub, 1981; Potter, 1975, 1976; Potter, Staub, Rado, & O'Connor, 2002). Here, we test the limits of viewers' detection ability by asking them to look for or recall named targets in sequences of six (Experiment 1) or twelve (Experiment 2) pictures they have never seen before, presented for durations between 13 and 80 ms per picture.

One reason for using such short durations is to investigate the possibility that the visual system has been configured by experience to process scene stimuli directly to an abstract conceptual level such as "a picnic." In feedforward computational models of the visual system (Serre, Kreiman, Kouh, Cadieu, Knoblich, & Poggio, 2007; Serre, Oliva, & Poggio, 2007) units that process a visual stimulus are hierarchically arranged: Units representing small regions of space (receptive fields) in the retina converge to represent larger and larger receptive fields and increasingly abstract information along a series of pathways from V1 to inferotemporal cortex

(IT) and higher to the prefrontal cortex. A lifetime of visual experience is thought to tune this hierarchical structure, which acts as a filter that permits categorization of objects and scenes with a single forward pass of processing. In this model, even a very brief, masked presentation might be sufficient for understanding a picture.

A widely accepted theory of vision, however, is that perception results from a combination of feedforward and feedback connections, with initial feedforward activation generating possible interpretations that are fed back and compared with lower levels of processing for confirmation, establishing reentrant loops (DiLollo, 2012; Di Lollo, Enns & Rensink, 2000; Enns & DiLollo, 2000; Hochstein & Ahissar, 2002; Lamme & Roelfsema, 2000). Such loops produce sustained activation that enhances memory. It has been proposed that we become consciously aware of what we are seeing only when such reentrant loops have been established (Lamme, 2006). A related suggestion is that consciousness arises from "recurrent long-distance interactions among distributed thalamo-cortical regions" (Del Cul, Baillet, & Dehaene, 2007, p. 2408). This network is ignited as reentrant loops in the visual system are formed (Dehaene, Kergsberg, & Changeux, 1998; Dehaene & Naccache, 2001; Dehaene, Sergent, & Changeux, 2003; see also Tononi & Koch, 2008). It has been estimated that reentrant loops connecting several levels in the visual system would take at least 50 ms to make a round trip, which would be consistent with stimulus onset asymmetries (SOAs) that typically produce backward masking

Thus, when people view stimuli for 50 ms or less with backward pattern masking, as in some conditions in the present study, there may be too little time for reentrant loops to be established between higher and lower levels of the visual hierarchy before earlier stages of processing are interrupted by the subsequent mask (Kovacs, Vogels & Orban, 1995; Macknik &

Martinez-Conde, 2007). In that case, successful perception would primarily result from the forward pass of neural activity from the retina through the visual system (Perrett, Hietanen, Oram, & Benson, 1992; Thorpe & Fabre-Thorpe, 2001; Hung, Kreiman, Poggio, & DiCarlo, 2005; DiCarlo, Zoccolan, & Rust, 2012). In support of the possibility of feedforward comprehension, Liu, Agam, Madsen, & Kreiman (2009) were able to decode object category information from human visual areas within 100 ms after stimulus presentation.

An open question is what level of understanding is achieved in the initial forward wave of processing. One behavioral approach to assessing how much is understood in the feedforward sweep is to measure the shortest time to make a discriminative response to a stimulus. Studies by Thorpe, Fabre-Thorpe and their colleagues (see reviews by Thorpe & Fabre-Thorpe, 2001, and Fabre-Thorpe, 2011) required subjects to make a go/no-go response to the presence of a category such as animals (or vehicles or faces) in photographs presented for 20 ms without a mask. They found that differential EEG activity for targets began at about 150 ms after presentation. The shortest above-chance reaction times (which would include motor response time) were under 300 ms. Choice saccades to a face in one of two pictures were even faster, as short as 100 ms (Crouzet, Kirchner, & Thorpe, 2010). In another study (Bacon-Mace, Kirchner, Fabre-Thorpe, & Thorpe, 2007), pictures with or without animals were followed by texture masks at SOAs between 6 and 107 ms; animal detection was at chance at 6 ms, but above-chance starting at 12 ms, with performance approaching an asymptote at 44 ms. These times suggested to the investigators that subjects were relying on feedforward activity, at least for their fastest responses.

A different question from that of feedback is whether a selective set or expectation can modify or resonate with the feedforward process (Llinas, Ribary, Contreras, & Pedroarena,

1998), obviating the need for reentrant processing and enabling conscious perception with presentation durations shorter than 50 ms. It is well known that advance information about a target improves detection. For example, in a recent study by Evans, Horowitz, and Wolfe (2011) participants viewed a picture for 20 ms preceded and followed by texture masks and judged whether they saw a specified target (e.g., animal, beach, street scene). Accuracy was consistently higher when the target was specified just before the stimulus was presented, rather than just after. Numerous studies have shown that selective attention has effects on the visual system in advance of expected stimuli (e.g., Cukur, Nishimoto, Huth, & Gallant, 2013). For example, in a recent study using multivoxel pattern analysis (Peelen & Kastner, 2011), the amount of preparatory activity in object-selective cortex that resembled activity when viewing objects in a given category was correlated with successful item detection.

To evaluate the effect of attentional set on target detection, in the present experiments we compared two conditions between groups. In one group the target was named just before the sequence (providing a specific attentional set) and in the other the target was named just after the sequence (with no advance attentional set). In the latter case, the participant had to rely on memory to detect the target. Previous studies of memory for pictures presented in rapid sequences have shown that memory is poor with durations in the range of 100-200 s per picture (Potter & Levy, 1969; Potter et al., 2002). Given these results and the known benefits of advance information already mentioned, we expected that advance information would improve performance; the question was whether it would interact with duration, such that detection of the target would be impossible at the shorter durations without advance information. Such a result would conflict with the hypothesis that feedforward processing, without reentrance and without top-down information, is sufficient for conscious identification.

While both the feedforward and feedback models predict that performance will improve with presentation time, the main questions addressed here are whether knowing the identity of the target ahead of time is necessary for successful detection, particularly at high rates of presentation, and whether there is a sharp discontinuity in performance as the duration of the images is reduced below 50 ms, as predicted by reentrant models. A seminal earlier study (Keysers, Xiao, Földiák, & Perrett, 2001) showed successful detection using RSVP at a duration as short as 14 ms, but the target was cued by showing the picture itself and pictures were reused for a single subject so that they became familiar. In the present experiments, by specifying the target with a name and by using multiple pictures in each sequence that the participants have never seen before, participants are forced to identify the target at a conceptual level rather than simply matching specific visual features.

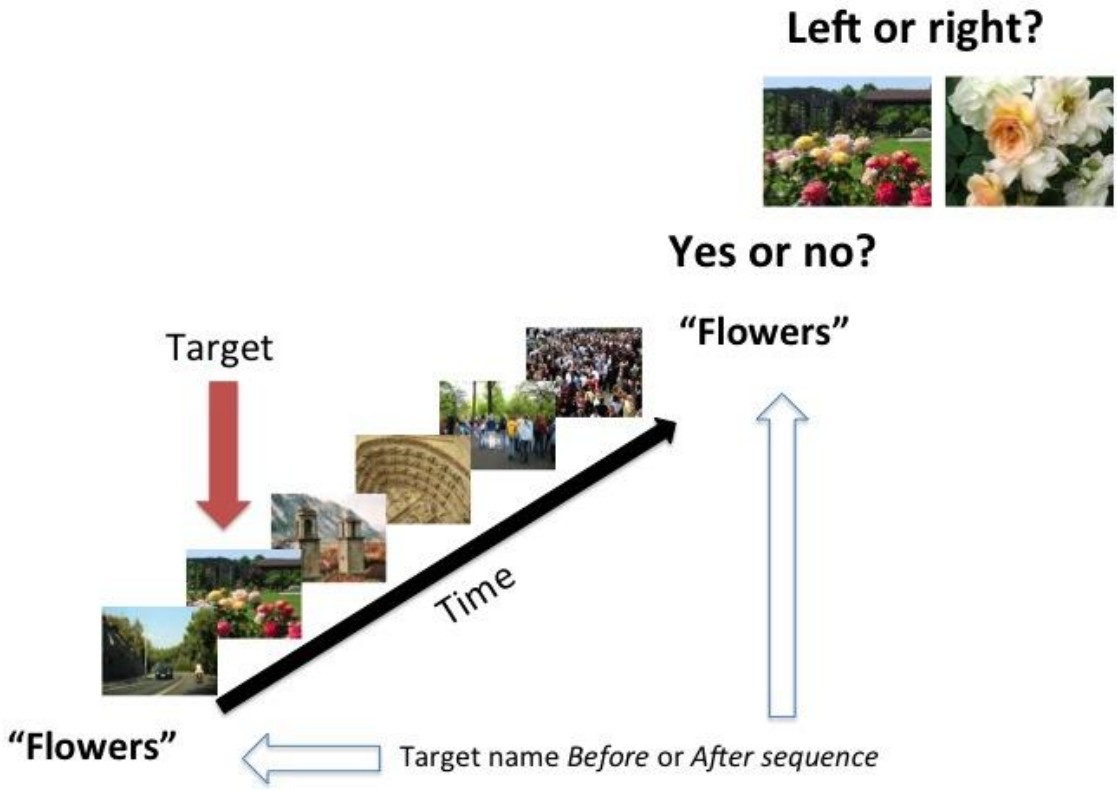

Figure 1. Illustration of a trial in Experiment 1. The target name appeared either 900 ms before the first picture or 200 ms after the last picture and the two forced-choice pictures appeared after the participant responded yes or no.

Experiment 1

Method

Procedure. Two groups of participants viewed an RSVP sequence of six pictures presented for 13, 27, 53, or 80 ms per picture and tried to detect a target specified by a written name. The one-to-four-word name reflected the gist of the picture as judged by the experimenters. Examples are: *swan, traffic jam, boxes of vegetables, children holding hands, boat out of water, campfire, bear catching fish,* and *narrow street*. For those in the Before group,

each trial began with a fixation cross for 500 ms, followed by the name for the target for 700 ms and then a blank of 200 ms and the sequence of pictures. A blank of 200 ms followed the sequence and then the question "Yes or No?" appeared and remained in view until the participant pressed Y or N on the keyboard to report whether he or she had seen the target. Those in the After group viewed a fixation cross for 500 ms at the beginning of the trial, followed by a blank of 200 ms and the sequence. At the end of the sequence there was a 200 ms blank and then the name was presented simultaneously with the yes/no question until the participant responded.

On trials in which the target had been presented, the participant's response was followed by a two-alternative forced choice between two pictures that matched the target name. The participant pressed the G or J key to indicate whether the left or right picture, respectively, had been presented. On no-target trials the words "No target" appeared instead of a pair of pictures.

Participants. The 32 participants (17 women) were volunteers 18-35 years of age. They were paid for their participation. All signed a consent form approved by the MIT Committee on the Use of Humans as Experimental Subjects. Participants were replaced if they made more than 50% false yes responses, overall, on nontarget trials, because such a high false alarm rate suggested that the participant was not following instructions, but randomly guessing. One participant was replaced in the Before group and three in the After group.

Materials. The stimuli were color photographs of scenes. The pictures were new to the participants and each picture was presented only once. For the targets, two pictures were selected that matched each target name; which one appeared in the sequence was determined randomly. The other picture was used as a foil in the forced-choice test after each target-present trial. The pictures were taken from the web and from other collections of pictures available for research use. They included a wide diversity of subject matter: indoor and outdoor, with and without

people. Pictures were resized to 300 x 200 pixels and were presented in the center of the monitor on a gray background. The horizontal visual angle was 10.3^{o} at the normal viewing distance of 50 cm. For the forced choice, two 300 x 200 pixel pictures were presented side by side.

Design. A practice block was presented at 133 ms per picture, followed by 8 experimental blocks of trials. Across blocks, the durations were 80, 53, 27, and 13 ms per picture, repeated in the next 4 blocks. There were 20 trials per block, including 5 no-target trials. The target, never the first or last picture, appeared in serial position 2, 3, 4, or 5, balanced over target trials within each block. Across every 8 participants, the 8 blocks of trials were rotated so that the pictures in each block of trials were seen equally often at each duration and in each half of the experiment.

Apparatus. The experiment was programmed with Matlab 7.5.0 and the Psychophysics Toolbox extension (Brainard, 1997) version 3, and was run on a Mac mini with 2.4 GHz, Intel Core 2 Duo processor. The Apple 17-in. CRT monitor was set to a 1024 x 768 resolution with a 75-Hz refresh rate. The room was normally illuminated. Timing errors sometimes occur in RSVP sequences (McKeeff, Remus, & Tong, 2007). Precision was controlled by using Wyble's Stream package for Matlab. We checked the actual timing on each refresh cycle in each of the groups and excluded trials in which there was a timing error of + /- 12 ms (equivalent to a single refresh of the monitor) or greater that affected the target picture or the pictures immediately before and after the target. As the timing errors were random, they increased noise in the data but did not produce any systematic bias. In Experiment 1 an average of 22% of the target trials were removed in the name-before group, and 10% in the name-after group. In Experiment 2 timing errors occurred on fewer than 1% of the trials.

Analyses. Repeated-measures analyses of variance (ANOVAs) were carried out on individual participants' d' measures as a function of before-after group and presentation duration (80, 53, 27, or 13 ms per picture). Planned paired *t* tests at each duration, separately for each group, compared d' with chance (0.0). Serial position effects were analyzed for the proportion of hits on target-present trials (as there was no way to estimate false yeses as a function of serial position, we did not use d'). Separate ANOVAs were carried out on the accuracy of the forced-choice responses on target-present trials, conditional on whether the participant had responded Yes (a hit) or No (a miss).

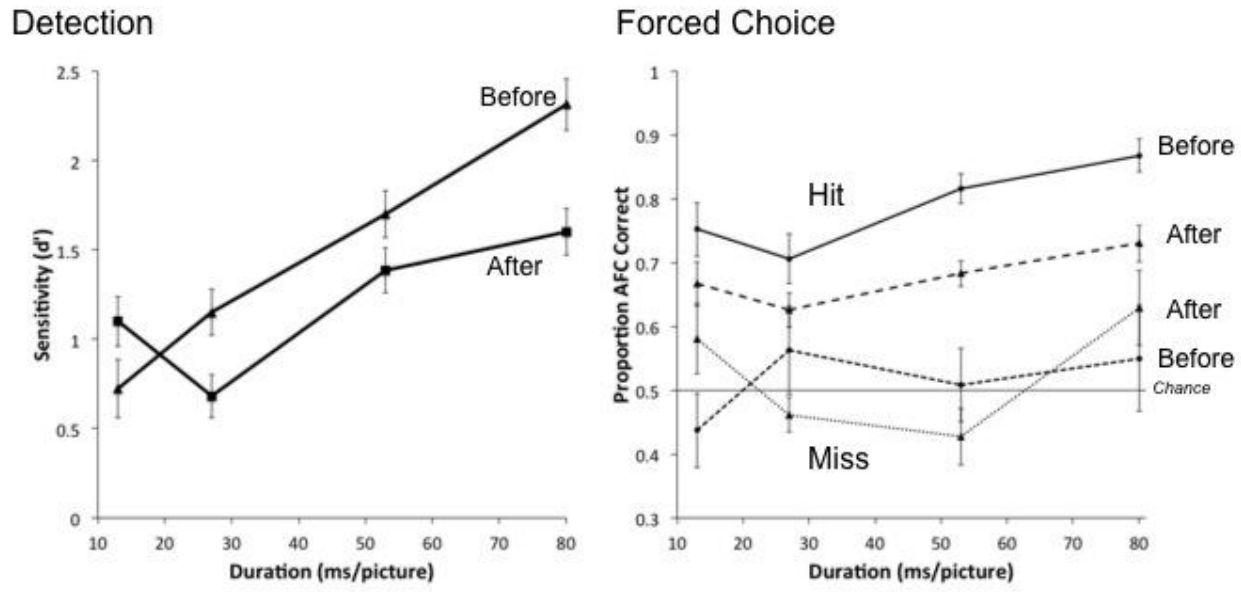

Figure 2. Results of Experiment 1 in which participants detected a picture that matched a name given before or after the sequence of six images (N = 16 in each group). Error bars depict the standard error of the means. **(A)** d' results of yes-no responses. (**B)** Proportion correct on two-alternative forced choice between two pictures with the same name on target-present trials, conditional on whether the participant had reported yes in the detection task (labeled “hit”) or no (“miss”). Chance = 0.5.

Results and Discussion

The results are shown in Figure 2. For the d' ANOVA of yes-no responses (Figure 2A), there were main effects of name position, $F(1, 30) = 4.792$, $p < .05$, $\eta_G^2 = .066$, and duration, $F(3, 90) = 38.03$, $p < .001$, $\eta_G^2 = .414$, and an interaction, $F(3, 90) = 7.942$, $p < .001$, $\eta_G^2 = .129$. As Figure 2 shows, having the target name presented before rather than after the sequence benefited detection substantially at 80 ms but not at all at 13 ms, with the other durations falling in between. Detection improved as the duration increased from 13 to 80 ms. Separate paired *t* tests, two-tailed, showed that d' was significantly above chance ($p < .001$) at each duration in each group. For the name-before group, at 13 ms, $t(15) = 4.45$, $p < .001$, SEM = 0.162; the significance of the difference increased at each of the other durations. For the name-after group, at 13 ms, $t(15) = 7.91$, $p < .0001$, SEM = 0.139; at 27 ms, $t(15) = 5.60$, $p < .0001$, SEM = 0.122; the significance of the difference increased at the other two durations,

In an ANOVA of the effect of serial position of the target on the proportion of correct yes responses, the main effect of serial position was significant, $F(3, 90) = 4.417$, $p < .01$, $\eta_G^2 = .023$. The means were .71, .71, .69, and .75 for serial positions 2, 3, 4, and 5, respectively, suggesting a small recency effect. A marginal interaction with name position, $F(3, 90) = 2.702$, $p = .05$, $\eta_G^2 = .014$, indicated that this effect was larger when the name came after the sequence. This was confirmed by separate analyses of serial position in the Before and After groups: serial position was not significant in the Before group ($p = .095$), but was significant in the After group, $F(3, 45) = 11.23$, $p < .001$, $\eta_G^2 = .073$, where the means were .67, .69, .67, and .75.

An ANOVA of the two-alternative forced-choice results on target-present trials (Figure 2B) showed that accuracy was high ($M = .73$) when participants had reported yes to the target (a hit) but at chance ($M = .52$) when they had reported no (a miss), $F(1, 30) = 57.92$, $p < .001$, $\eta_G^2 = .253$. The main effect of group (Before/After) was significant, $F(1, 30) = 6.70$, $p < .05$, η_G^2

=.018, and interacted with whether the response had been yes or no, $F(1, 30) = 4.63$, $p < .05$, η_G^2 =.026. When participants reported having seen the target, forced choice accuracy was relatively better in the Before condition than the After condition, although both were above chance. When the target was missed, both groups were close to chance. There was a main effect of duration, $F(3, 90) = 3.76$, $p < .05$, η_G^2 =.048, and no other significant interactions.

The main findings of Experiment 1 are that viewers can detect and retain information about named targets they have never seen before at an RSVP duration as short as 13 ms, and that they can do so even when they have no information about the target until after presentation. Furthermore, there was no clear discontinuity in performance as duration was decreased from 80 to 13 ms. If reentrant feedback from higher to lower levels played a necessary role in extracting conceptual information from an image, one would expect an inability to detect any targets at 27 and 13 ms even in the Before condition, contrary to what we observed. If advance information about the target resonated or interacted with incoming stimuli, accounting for successful performance at 27 and 13 ms without reentrance, then performance should have been near chance at those durations in the After condition, again contrary to the results. A feedforward account of detection is more consistent with the results, suggesting that a presentation as short as 13 ms, even when masked by following pictures, is sufficient on some trials for feedforward activation to reach a conceptual level without selective attention.

Experiment 2

One question about the results of Experiment 1 is whether they would generalize to sequences longer than six pictures. Given that targets were limited to only four serial positions (excluding the first and last picture), could participants have focused on just those pictures, maintaining one or more of them in working memory to compare subsequently to the target

name? In that case, increasing the number of pictures to 12 (in Experiment 2) should markedly reduce the proportion detected, at least in the name-after condition.

Method

The method was the same as that of Experiment 1 except as noted.

Participants. The 32 participants (22 women) were volunteers of 18-35 years of age, most of them college students; none had participated in Experiment 1. They were paid for their participation. All signed a consent form approved by the MIT Committee on the Use of Humans as Experimental Subjects. Participants were replaced if they made more than 50% false yes responses, overall, on nontarget trials. No participant was replaced in the Before group; four were replaced in the After group.[1]

Design. The design was like that of Experiment 1, with two groups of participants, one with the target presented before the sequence, the other with the target presented after the sequence. The main difference was that trials consisted of 12 rather than 6 pictures. To make the 12-picture sequences, two 6-picture sequences from Experiment 1 were combined by randomly pairing trials in a given block, with the restriction that the two targets in a pair were in the same serial positions (2, 3, 4, or 5; after combination, the two potential targets were in serial positions 2 and 8, or 3 and 9, etc.). To end up with an even number of 6-item sequences, we generated two new 6-picture trials per block, one with a target and one without. There were 11 trials per block, 8 with targets and 3 without. Each of the eight target serial positions was represented once per block. Which of the two target names was used was counterbalanced between subjects within group. Across participants within group, the 8 blocks of trials were rotated so that the pictures in each block of trials were seen equally often at each duration and in each half of the experiment.

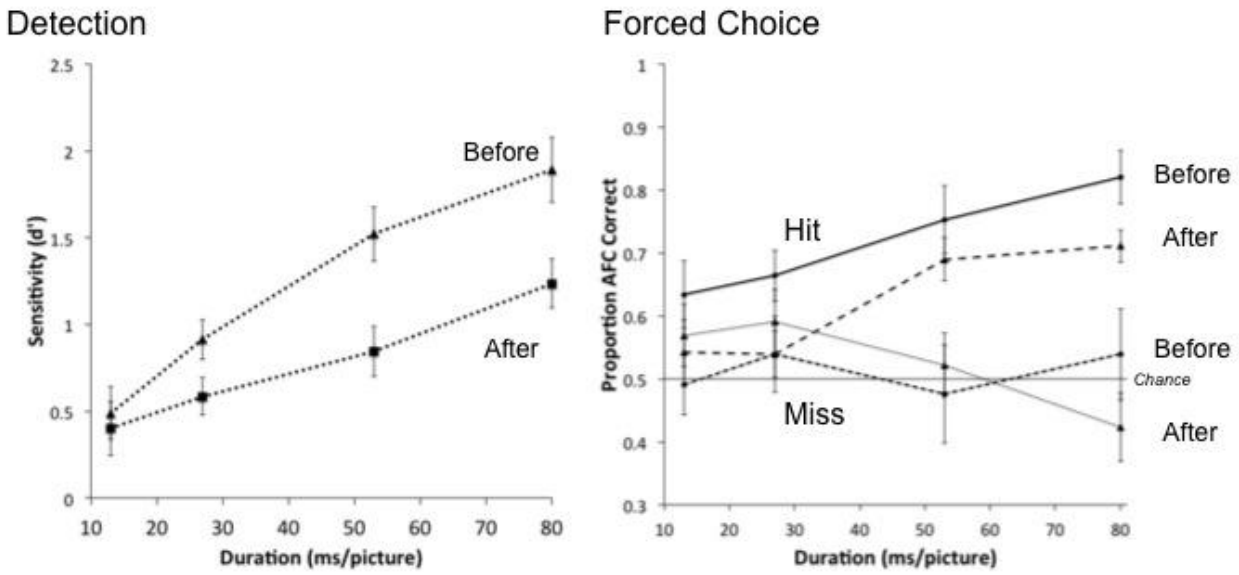

Figure 3. Results of Experiment 2 in which participants detected a picture that matched a name given before or after the sequence of 12 images (N = 16 in each group). Error bars depict the standard error of the means. **(A)** d' results of yes-no responses. (**B)** Proportion correct on two-alternative forced choice between two pictures with the same name on target-present trials, conditional on whether the participant had reported yes in the detection task (labeled "hit") or no ("miss"). Chance = 0.5.

Results and Discussion

The results are shown in Figure 3. The main results were similar to those of Experiment 1. In the d' analysis of the yes-no responses, there were main effects of whether the name was given before or after, $F(1, 30) = 8.785$, $p < .01$, $\eta_G^2 = .083$, and duration, $F(3, 90) = 28.67$, $p < .001$, $\eta_G^2 = .397$. There was more accurate detection when the name was given before the sequence rather than after, and detection improved as the duration increased from 13 to 80 ms. The interaction was not significant ($p = .22$). Separate paired *t* tests, two-tailed, showed that d' was significantly above chance ($p < .02$) at each duration in each group. For the name-before group, at 13 ms, $t(15) = 3.28$, $p < .01$, SEM = 0.152; the significance of the difference increased

at each of the other durations. For the name-after group, at 13 ms, $t(15) = 2.83$, $p < .02$, SEM = 0.155; the significance of the difference increased at each of the other durations.

In an ANOVA of the effect of serial position of the target on the proportion of correct yes responses, the main effect of serial position was significant, $F(7, 210) = 5.20$, $p < .001$, $\eta_G^2 = .032$. The means were .57, .54, .66, .76, .66, .63, .64, and .62 for serial positions 2, 3, 4, 5, 8, 9, 10, and 11, respectively, suggesting a slight disadvantage for primacy, but no recency benefit. There were no interactions.

An ANOVA of the two-alternative forced-choice results on target-present trials (Figure 2B) showed that accuracy was fairly high ($M = .67$) when participants had reported yes to the target (a hit) but near chance ($M = .52$) when they had reported no (a miss), $F(1, 30) = 20.61$, $p < .001$, $\eta_G^2 = .122$. The main effect of group (Before/After) was not significant, $F(1, 30) = 2.34$, $p = .136$, $\eta_G^2 = .018$, but there was a marginal interaction with whether the response had been yes or no, $F(1, 30) = 2,88$, $p = .10$, $\eta_G^2 = .019$. As in Experiment 1, having the name before was only better than having the name after when the participant reported having seen the target; when the trial was a miss, both groups were close to chance. There was no main effect of duration, $F(3, 90) = 1.35$, but there was an interaction with hit/miss, $F(3, 90) = 6.43$ $p < .01$, $\eta_G^2 = .064$. As seen in Figure 3B, the hit benefit was larger at longer durations.

Altogether, the results of Experiment 2 replicated the main results of Experiment 1, but now with 12 pictures per sequence rather than 6 (see Figure 4). An ANOVA compared the d' results of the two experiments. Performance (d') was significantly higher with 6-picture sequences (M = 1.33) than with 12-picture sequences (M = 1.06), $F(1, 60) = 9.83$, $p < .01$, $\eta_G^2 = .057$. No interactions with experiment were significant.

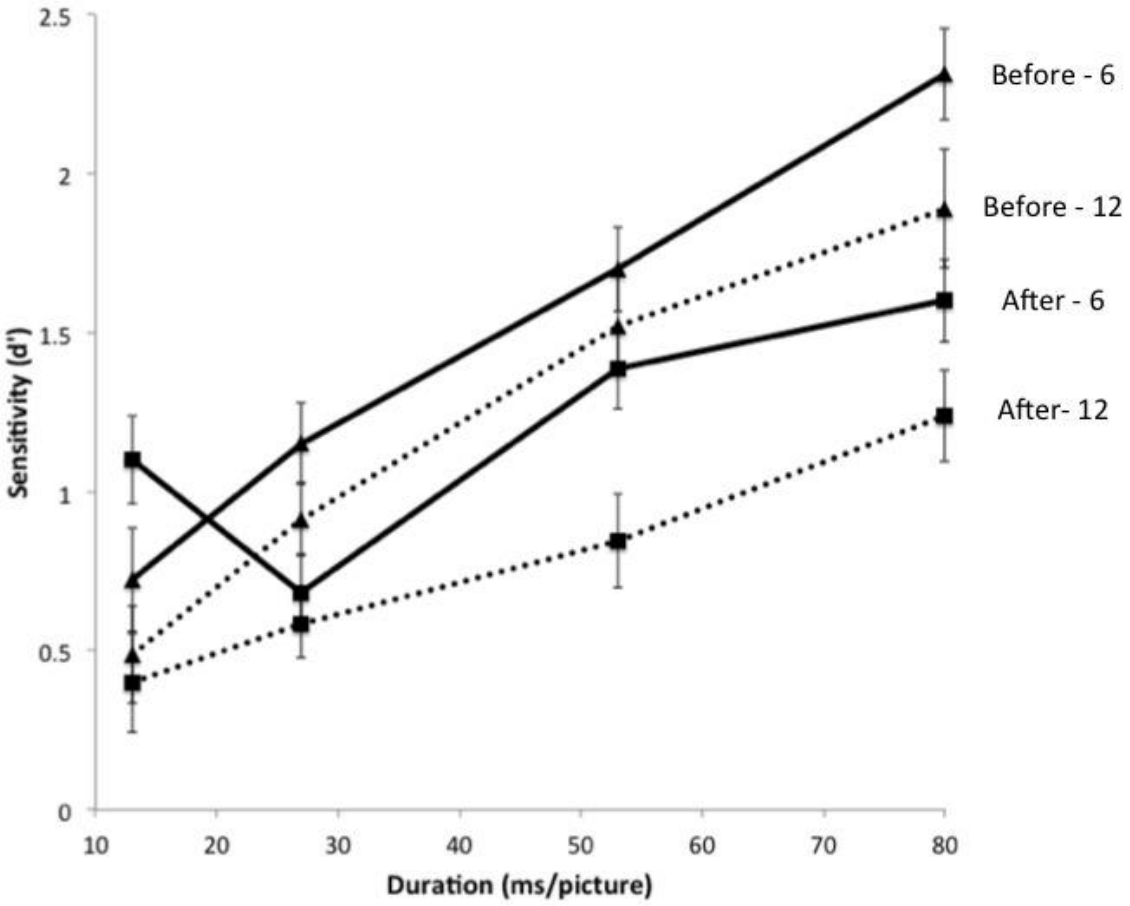

Figure 4. A comparison of the d' results of Experiment 1 (6 pictures) and Experiment 2 (12 pictures). Error bars depict the standard error of the means.

Clearly, we can reject the hypothesis that participants could encode only two or three pictures in working memory; otherwise performance would have fallen more dramatically in Experiment 2, especially in the After condition in which participants had to retain information about the pictures for later retrieval.

The results also demonstrate that a specific attentional expectation is not required for extracting conceptual information from a stream of pictures: performance remained substantially above chance at all durations when the target was specified after the sequence. The forced-choice results indicate, however, that visual details were lost at the two shorter durations with 12

pictures to retain, even when the target was correctly reported. In the After condition, however, the forced-choice test was slightly delayed relative to the Before condition, because the participants had to read the name of the target and scan their memory of the sequence before making a yes or no response. This intervening processing may account for the somewhat reduced performance in the forced-choice task in both Experiments 1 and 2, when the target name followed the sequence.

General Discussion

The results of both experiments show that conceptual understanding can be achieved when a novel picture is presented as briefly as 13 ms and masked by other pictures. Even when participants were not given the target name until after they viewed the entire sequence of six or twelve pictures, their performance was above chance even at 13 ms, indicating that a top-down attentional set is not required to rapidly extract and at least briefly retain conceptual content from an RSVP stream. The number of pictures in the sequence and their serial positions had little effect on performance, suggesting that pictures were processed immediately rather than accumulating in a limited-capacity memory buffer for subsequent processing. This pattern of results supports the hypothesis that feedforward processing is sufficient for conceptual comprehension of pictures.

As expected, detection was more accurate the longer the duration per picture. However, it was striking that there was no sharp drop in detection at or below a duration of 50 ms, contrary to predictions of feedback or reentrant models of conscious perception (e.g., Del Cul et al., 2007; Lamme, 2006). Instead, performance declined gradually with shorter durations but remained well above chance at 13 ms. Moreover, when viewers reported that they had detected the target, they were usually above chance in selecting it in a forced choice between two pictures, both of which

fit the target name. That is, they remembered more about the picture than simply the concept provided by the target name. When they had not detected the target their forced choice was near chance, suggesting that visual features of unidentified pictures were not retained.

Although the present behavioral results cannot entirely rule out feedback, they do present a challenge to existing reentrant models. They also raise a further question: How can conceptual understanding persist long enough to be matched to a name presented 200 ms after the offset of the final, masking picture, given that the target might have been any of the 6 or 12 pictures just viewed? The answer to this question may lie in the carwash metaphor of visual processing (Moore & Wolfe, 2001), in which each stimulus is passed from one level of processing to the next. In such a model, multiple stimuli can be in the processing pipeline at once. At the end of this pipeline, the stimuli, having now been processed to the level of concept, may persist in local recurrent networks that sustain activation for several pictures in parallel, at least briefly. In such a model, concepts are presumably represented in a multi-dimensional, sparsely populated network in which visual masks may not be effective if they are not also conceptually similar to the item being masked. The finding that a forced choice between two conceptually equivalent pictures is above chance only if the participant correctly detected the target is consistent with the conjecture that when feedforward processing does not reach a conceptual level, lower levels of representation are already masked and no featural information can be accessed.

The finding that observers can perceive and comprehend conceptual information from such brief images extends previous evidence that a purely feedforward mode of processing is capable of decoding complex information in a novel image (e.g., DiCarlo et al., 2012; Serre, Kreiman et al., 2007; Thorpe, Fize, & Marlot, 1996). Feedforward models are consistent with a range of neural results. For example, in a study by Keysers et al. (2001), recordings were made

of individual neurons in the cortex of the anterior superior temporal sulcus (STSa) of monkeys as they viewed continuous RSVP sequences of pictures; the monkey's only task was to fixate on the screen. Neurons in STSa that were shown to respond selectively to a given picture at a relatively slow presentation rate of 200 ms per picture also responded selectively (although not as strongly) to the same picture at presentations as short as 14 ms per image.

The present behavioral results suggest that feedforward processing is capable of activating the conceptual identity of a picture even when reentrant processing has presumably been blocked because the picture is briefly presented and is then masked by immediately following pictures. As participants were capable of reporting the presence of a target under these conditions, the results strongly suggest that reentrant processing is not always necessary for conscious processing. They are consistent with the possibility, however, that reentrant loops facilitate processing and may be essential to comprehend details of the image. For example, a rapid but coarse first pass of low spatial frequency information may provide global category information that is subsequently refined by reentrant processing (e.g., Bar, Kassam, Ghuman, et al., 2006). Work with monkeys has found that neurons that are selective for particular faces at a latency of about 90 ms give additional information about facial features beginning about 50 ms later (Sugase, Yamane, Ueno, & Kawano, 1999). Reentrant processing therefore might be involved after an initial feedforward pass (DiLollo, 2012).

The present findings can be contrasted with those of masked priming studies in which the prime is not consciously seen although it has an effect on the response to an immediately following stimulus. In a typical experiment a brief presentation of a word in the range of 25-60 ms – the prime – is followed by a second unmasked word to which the participant must respond (Dehaene, Naccache, Cohen, Le Bihan, Mangin, Poline, & Riviere, 2001; Forster & Davis,

1984). If the prime is identical to or is related to the target word, the response to the latter is faster and more accurate than without the prime or with an unrelated prime, even when the prime is not consciously identified. In such studies the participant's focus of attention is on the final word, whose longer duration permits it to receive full, recurrent processing that may block awareness of the more vulnerable information from the prime that was extracted during the feedforward sweep. In the present experiments, in contrast, the masking stimuli are the same duration as the preceding target stimulus and all are potential targets. In these conditions, even durations as short as 13 ms are clearly sufficient, on a significant proportion of trials, to drive conscious detection, identification, and immediate recognition memory.

Finally, perhaps our most striking finding is that performance was surprisingly good even when the target name was given only after the sequence. It has long been assumed that detection of rapidly presented targets in an RSVP stream (e.g., Potter, 1976) is possible only because the participants had the opportunity to configure their attentional filters in advance (e.g., Peelen & Kastner, 2011). Indeed, Potter (1976) found that the ability to detect an image named in advance was much greater than the ability to recognize pictures later in a yes-no test of all the pictures mixed with distractors. Other research (e.g., Potter et al., 2002) indicates that memory traces generated by RSVP are fragile and are rapidly degraded by successive recognition testing. When participants are given an immediate recognition test of just one item from the stream, however, the present results show that they are able to detect it in their decaying memory trace at a level of accuracy not far from accuracy when the target was pre-specified at the start of the trial. This result is consistent with the idea that a single forward sweep as short as 13 ms is capable of extracting a picture's conceptual meaning without advance knowledge. Moreover, the pictures'

conceptual identities can be maintained briefly, enabling one to be matched to a name presented after the sequence.

A possible role for such rapid visual understanding in normal experience is to provide nearly instantaneous conceptual activation that enables immediate action when necessary, without waiting to refine understanding by reentrant processing or by the kind of conscious reflection that requires a stable recurrent network.

References

Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Dale, A.M., Ha¨ma¨ la¨ inen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences USA*, *103*, 449–454.

Bacon-Mace, N., Kirchner, H., Fabre-Thorpe, M., & Thorpe, Simon. J. (2007). Effects of task requirements on rapid natural scene processing: From common sensory encoding to distinct decisional mechanisms. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1013-1026.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision, 10* (4):16, 1–17.

Cukur, T., Nishimoto, S., Huth, A. G, , & Gallant, J. I. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience, 16* (6), 763-770.

Dehaene, S., Kergsberg, M.. & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences USA*, 95:14529--14534.

Dehaene, S. & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition, 79*, 1–37.

Dehaene, S., Naccache, L., Cohen, L., LeBihan, D., Mangin, J. F., Poline, J.-B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, 4:752--758.

Dehaene, S., Sergent, C., & Changeux, J.-P., (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences USA, 100 (14),* 8520–8525.

Del Cul, A., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology 5*, 2408-2423.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012).How does the brain solve visual object recognition? *Neuron, 73*, 415-434. DOI 10.1016/j.neuron.2012.01.010

DiLollo, V. (2012).The feature-binding problem is an ill-posed problem. *Trends in Cognitive Sciences, 16* (6), 317-321.

Di Lollo, V., Enns, J. T., Rensink, R. A. (2000). Competition for consciousness among visual events: the psychophysics of reentrant visual pathways. *Journal of Experimental Psychology: General, 129*, 481-507.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences, 4*, 345-352.

Evans, K. K., Horowitz, T. S., & Wolfe, J. W. (2011). When categories collide: Accumulation of information about multiple categories in rapid scene perception. *Psychological Science, 22*, 739–746. (2011)

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Psychology, 2: 243*. doi: 10.3389/fpsyg.2011.00243

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 680-698.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36* (5): 791-804.

Hung, C.P., Kreiman, G., Poggio, T., & DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310*, 863-866.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 604-610.

Keysers, C., Xiao, D. K., Földiák, P., & Perrett, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience, 13*, 90–101.

Keysers, C., Xiao, D.-K., Földiák, P., Perrett, D.I. (2005). Out of sight but not out of mind: The neurophysiology of iconic memory in the superior temporal sulcus. *Cognitive Neuropsychology 22*, 316-332.

Kovacs, G., Vogels, R., & Orban, G. A. (1995). Cortical correlate of backward masking. *Proceedings of the National Academy of Sciences U.S.A., 92*, 5587-5591.

Lamme. V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10* (11), 494-501,

Lamme. V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neuroscience, 23*, 571-579.

Liu, H. , Agam, Y. Madsen, J. R., Kreiman, G. (2009). Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron***.** *62*, 281-290.

Llinas, R., Ribary, U., Contreras, D. & Pedroarena, C. (1998). The neuronal basis for consciousness. *Philos. Trans. R. Soc. London B, 353*, 1841–1849.

Loschky, L. C., Hansen, B.C., Sethi, A., & Pydimarri, T. N. (2010). The role of higher order image statistics in masking scene gist recognition. *Attention, Perception, & Psychophysics, 72*(2), 427-444.

Macknik, S. L., & Martinez-Conde, S. (2007). The role of feedback in visual masking and visual processing. *Advances in Cognitive Psychology, 3*, 125-152.

McKeeff, T.J., Remus, D.A., & Tong, F. (2007). Temporal limitations in object processing across the human ventral visual pathway. *Journal of Neurophysiology, 98*, 382-393.

Moore, C. M., & Wolfe, J. M. (2001). Getting beyond the serial/parallel debate in visual search: A hybrid approach. In K. Shapiro (Ed.) *The limits of attention: Temporal constraints on human information processing*. (pp. 178-198) Oxford: Oxford University Press.

Peelen, M. V., & Kastner, S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex.*Proceedings of the National Academy of Sciences USA, 108 (29),* 12125-12130.

Perrett, D., Hietanen, J., Oram, M., & Benson, P. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London Series B. 335*, 23–30.

Potter, M.C. (1975). Meaning in visual search. *Science, 187,* 965-966.

Potter, M.C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 509-522.

Potter, M. C., Staub, A., Rado, J., & O'Connor, D. H. (2002). Recognition memory for briefly-presented pictures: The time course of rapid forgetting. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 1163-1175.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research, 165,* 33-56.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, 104,* 6424–6429.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature, 400,* 869–873.

Thorpe, S. & Fabre-Thorpe, M. (2001). Seeking categories in the brain. *Science, 291,* 260–263.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520-522.

Tononi, G., & Koch, C. (2008). The neural correlates of consciousness: An update. *Annals of the New York Academy of Science, 1124,* 239-261.

M.C.P. developed the study concept. All authors contributed to the study design. Testing, data collection, and data analysis were performed by E.S.M. and C.E.H. under the supervision of M.C.P. and B.W. M.C.P. drafted the paper and B.W. and C.E.H. provided critical revisions. All authors approved the final version of the paper for submission.

Acknowledgements

This work was supported by Grant MH47432 from NIMH. We thank Chidinma Egbukichi for assistance.

Footnote

1. Because of the relatively large number of replaced subjects in Experiment 2's After group, we also ran the main d' analysis with the original 16 subjects. Although d' was slightly lower with the original group than with the replaced subjects, none of the significance levels changed, including the comparison with the Before group.

Figure Captions

Figure 1. Illustration of a target-present trial in Experiment 1's target-before condition. The target name appeared 900 ms before the first picture, the question "Yes or No?" appeared after the final picture, and the two forced-choice pictures appeared after the participant responded yes or no.

Figure 2. Results of Experiment 1 in which participants detected a picture that matched a name given before or after the sequence of six images (N = 16 in each group). Error bars depict the standard error of the means. **(A)** Proportion yes responses to target (hits) and non-target trials (false alarms). (**B)** Proportion correct on two-alternative forced choice between two pictures with the same name on target-present trials, conditional on whether the participant had reported yes in the detection task (labeled “hit”) or no (“miss”). Chance = 0.5.

Figure 3. Results of Experiment 2 in which participants detected a picture that matched a name given before or after the sequence of 12 images (N = 16 in each group). Error bars depict the standard error of the means. **(A)** Proportion yes responses to target (hits) and non-target trials (false alarms). (**B)** Proportion correct on two-alternative forced choice between two pictures with the same name on target-present trials, conditional on whether the participant had reported yes in the detection task (labeled “hit”) or no (“miss”). Chance = 0.5.

Figure 4, A comparison of the d' results of Experiment 1 (6 pictures) and Experiment 2 (12 pictures). Error bars depict the standard error of the means.