# Rapid and accurate processing of multiple objects in briefly presented scenes

**Henry Railo** Department of Psychology, Centre for Cognitive Neuroscience, and Turku Brain and Mind Centre, University of Turku, Turku, Finland

**Veli-Matti Karhu** Department of Psychology, University of Turku, Turku, Finland

**Jeremy Mast** Department of Psychology, University of Turku, Turku, Finland

**Henri Pesonen** Turku Brain and Mind Centre and Department of Mathematics and Statistics, University of Turku, Turku, Finland

**Mika Koivisto** Department of Psychology, Centre for Cognitive Neuroscience, and Turku Brain and Mind Centre, University of Turku, Turku, Finland

**Humans can detect multiple objects in briefly presented natural visual scenes, but the mechanisms through which the objects are segmented from the background and consciously accessed remain open. By asking participants to report how many humans natural photos presented for 50 ms contain, we show that up to three items can be rapidly enumerated from natural scenes without compromising speed or accuracy. In contrast to standard parallel and serial models of object selection, our results revealed that the participants were fastest in enumerating two objects; even enumerating one single item required additional processing time. Also enumeration accuracy slightly increased in the subitizing range as number increased. Our results suggest that the visual system is tuned to process multiple items, which may underlie spatial and numerical cognition, and be beneficial in real-world situations that often require dealing with more than one object at a time.**

## Introduction

In everyday life humans operate in situations that require them to process multiple items. Can the brain process multiple stimuli simultaneously, or is information about different objects integrated serially (e.g., Cavanagh & Alvarez, 2005; Wolfe, Oliva, Horowitz, Butcher, & Bombas, 2002)? Since the 19th century, researchers have probed this question by asking participants to report the number of visually presented objects as fast as possible (Jevons, 1871). The results have revealed a process termed *subitizing* through which humans can enumerate small sets of objects (up to three to four) fast and extremely accurately without counting (Kaufman, Lord, Reese, & Volkman, 1949; Piazza, Fumarola, Chinello, & Melcher, 2011; Revkin, Piazza, Izard, Cohen, & Dehaene, 2008). Subitizing is assumed to reflect a fundamental capacity-limited process that enables humans to deal with multiple objects in the world (Cavanagh & Alvarez, 2005; Cowan, 2001; Pylyshyn, 2001). This ability has also been proposed to play a central role in the development of numerical cognition (Feigenson, Dehaene, & Spelke, 2004). Previous enumeration studies have used simple geometric shapes presented on a uniform background as stimuli, and one of the aims of the present study was to verify that objects that are embedded in real-word contexts can be efficiently enumerated through subitizing. The second aim was to examine how efficiently humans process real-world scenes that contain different

Citation: Railo, H., Karhu, V.-M., Mast, J., Pesonen, H., & Koivisto, M. (2016). Rapid and accurate processing of multiple objects in briefly presented scenes. *Journal of Vision*, *16*(3):8, 1–11, doi:10.1167/16.3.8.

doi: 10.1167/16.3.8 Received June 4, 2015; published February 5, 2016 ISSN 1534-7362

numbers of target objects. How fast are multiple objects segmented from the background? Does an increase in the number of objects to be processed always increase processing times?

The phenomenon of subitizing suggests that humans can process a small number of visual items without eliciting additional costs on information processing. This is revealed through a characteristic two-segment performance curve when enumeration performance is plotted as a function of number of items: For the first few items, referred to as the subitizing range, enumeration performance remains approximately constant, after which reaction times (RTs) increase and accuracies decrease (Trick & Pylyshyn, 1994). Although earlier theories assumed that subitizing reflects parallel, preattentive visual processing (Trick & Pylyshyn, 1994), later studies have demonstrated that it is dependent on visual attention (Railo, Koivisto, Revonsuo, & Hannula, 2008; Vetter, Butterworth, & Bahrami, 2008). When the items to be enumerated cannot easily be resolved from other items or the background in enumeration studies, subitizing performance suffers (Trick & Pylyshyn, 1994; Watson, Maylor, Allen, & Bruce, 2007), suggesting that items in real-word scenes may be more challenging to enumerate than simple geometric shapes. Subitizing and similar results in multiple object-tracking paradigms (Alvarez & Cavanagh, 2005; Pylyshyn & Storm, 1988) have led researchers to propose that selective visual attention can select multiple objects simultaneously (Cavanagh & Alvarez, 2005; Huang & Pashler, 2007). Also serial attention switching has been proposed to contribute to processing multiple objects (Oksama & Hyönä, 2008).

Humans can rapidly analyze the gist of briefly presented natural scenes despite the apparent complexity of the task (Fei-Fei, Iyer, Koch, & Perona, 2007). For instance, participants can rapidly classify natural images according to whether they contain targets from prespecified categories or not (Thorpe, Fize, & Marlot, 1996), even when attention is allocated to another task (Cohen, Alvarez, & Nakayama, 2011; Li, Van Rullen, Koch, & Perona, 2002). However, when the scene to be categorized contains four foreground objects, categorization performance suffers under dual-task conditions, suggesting that processing complex scenes that contain multiple objects requires serial attentional processing (Walker, Stafford, & Davis, 2008). Yet, participants can report if one animal is present in one of two scenes as fast as they report the presence of an animal in a single image (Fei-Fei, VanRullen, Koch, & Perona, 2005; Rousselet, Fabre-Thorpe, & Thorpe, 2002), provided that the two scenes are presented sufficiently far away from each other to minimize interference (VanRullen, Reddy, & Fei-Fei, 2005). Other studies have reported that behavioral performance (RTs and accuracy) suffers if the search for a target object has to be performed simultaneously on multiple scenes (Rousselet, Thorpe, & Fabre-Thorpe, 2004a, 2004b). As stated, this decrease in performance is partly explained by interstimulus spacing (VanRullen et al., 2005, but see Fei-Fei et al., 2005), but it could also be due to the fact that processing multiple different scenes is extremely demanding and artificial (Rousselet et al., 2004b). In real life, humans are typically required to resolve objects embedded in a single scene. VanRullen and Koch (2003) showed that participants correctly reported two to three items from briefly presented scenes each of which contained 10 different objects. However, the result leaves open the question whether processing multiple objects from natural scenes is associated with costs in processing time relative to a single stimulus condition (e.g., due to serial shifts of attention).

To directly test how behavioral performance changes as the number of objects to be processed increases, and whether subitizing generalizes to real-world scenes, we asked participants to enumerate how many humans natural photos contain. As shown in Figure 1, a condition where the background was replaced with a uniform color was used as a baseline enumeration condition. As this resembles typical subitizing studies, the uniform background condition was assumed to yield the characteristic two-segment performance curve where enumeration performance remains roughly constant for small numbers of objects. A similar performance curve should be observed in the full scene condition if a small number of objects can be efficiently detected and enumerated also from naturalistic images. If subitizing does not generalize to naturalistic scenes—for example, because detecting multiple humans from natural scenes requires serial shifts of attention—enumeration performance should decrease as the function of number already with small number of humans.

# Methods

## Participants

After pilot testing (three participants), we set out to test 30 participants in order to have sufficient statistical power to detect relatively small variations in the RTs in the subitizing range (e.g., Railo, 2014). Thirty-three undergraduate students (25 females, 18–36 years old, median age 22) took part in the experiment, and received study credits for participation. The participants reported normal or corrected-to-normal vision. Informed written consent was obtained before the

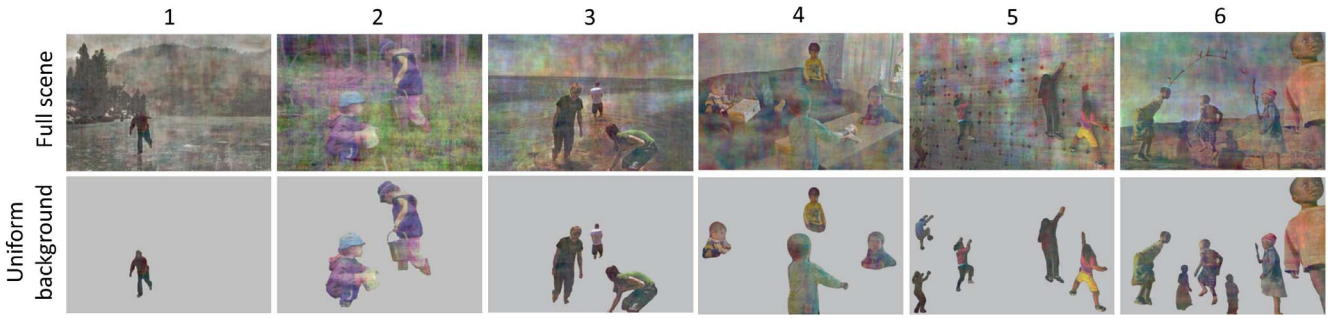


Figure 1. Examples of the stimuli used in the experiment. The number of humans in the photos varied between one and six, and in the baseline condition the background was replaced with a uniform gray color. Low-level properties (luminance histograms and Fourier amplitude spectra: spatial frequency, orientation, amplitude, and phase) of the images were equated to minimize the impact of low-level stimulus confounds on performance differences between number conditions (Willenbockel et al., 2010).

experiment, and the study was conducted in accordance with the Declaration of Helsinki.

## Stimuli and procedure

Participants were presented with color images that contained one to six humans in various real-world settings, either indoors or outdoors. We used humans as target items as they are an important visual object category in real-world settings (Fei-Fei et al., 2007), and familiar to all participants. The target stimuli were acquired by photographing and from the Internet. The stimulus set contained 168 photos in total. Each number condition (1–6) contained 28 images. In the photos, humans were typically distributed around the image, they never overlapped with each other, and at least their head and shoulders were visible. Because the stimuli depicted real-world settings, they sometimes contained also other objects than humans (e.g., chairs, cars). To construct control images that mimic traditional subitizing studies, the background was replaced with a uniform gray color ($\sim$68 cd/m$^2$). Thus, the control condition included exactly the same items as the full scene condition, but with uniform backgrounds (Figure 1).

Low-level statistical properties of the images were equated with Matlab 7.9.0 (The MathWorks, Inc., Natick, MA) using the SHINE toolbox by matching luminance histograms (histMatch function) and Fourier amplitude spectra (specMatch function, which matches spatial frequency, orientation, amplitude and phase; Willenbockel et al. 2010). This matching was performed in an iterative manner (10 repetitions) to jointly match both luminance histograms and Fourier amplitude spectra (Willenbockel et al., 2010). The RGB layers of the images were matched separately, and the processed layers were recombined to produce color images. This ensures that changes in enumeration performance between different number conditions (in the full scene condition) are likely not due to low-level image properties but are related to the top-down effects of enumeration. The images with uniform backgrounds were constructed after matching the low-level properties. Mean luminance of the full scene stimuli was $\sim$40 cd/m$^2$.

Each trial began with the presentation of a fixation point on a light gray background ($\sim$110 cd/m$^2$) for 1 s, after which a target image (900 $\times$ 600 pixels, $\sim$23°) was displayed for 50 ms. The stimuli were presented on a 21-in. CRT-monitor set at 60 Hz (1024 $\times$ 768 resolution), with stimulus presentation and data collection controlled by E-Prime 1.2 software (Psychology Software Tools, Pittsburgh, PA). The participants reported the number of humans as fast as possible by speaking the number word to a microphone (AKG D40S, AKG Acoustics, Vienna, Austria). RTs were measured using a voice key (Psychology Software Tools, model 200A). After this the participant logged the response by pressing the corresponding number key.

The full scene and uniform background conditions were completed in different blocks (84 stimuli each) that contained different images (presentation order counterbalanced). Each participant saw each stimulus once (eliminating stimulus-specific learning effects), and across participants each stimulus appeared equally often in the full scene and uniform background conditions. To control for possible differences in pronouncing different number words, a condition where participants named number symbols (1–6) presented in the center of a monitor in randomized order was included to the experiment (speeded responses; 15 trials/number). This number-naming control condition was always conducted last. Each participant completed 10 practice trials before the enumeration and number-naming conditions (the

| Response word | Median RT (ms) | 95% CI |
|---|---|---|
| One | 307.8 | [296.5, 318.3] |
| Two | 305.3 | [295.1, 314.1] |
| Three | 310.3 | [301.1, 320.3] |
| Four | 335.8 | [320.9, 356.0] |
| Five | 334.2 | [311.8, 353.8] |
| Six | 328.9 | [315.6, 343.5] |

Table 1. RTs in the number-naming control condition.

images presented during the practice trials were not presented during the actual experiment).

## Statistical analysis

Data were analyzed in R statistical software (version 3.1.2; R Development Core Team, 2014). Before statistical analysis of RTs, median RTs in the number-naming control condition (shown in Table 1 with 95% confidence intervals [CIs] based on 1,000 bootstrap samples) were subtracted from the single-trial RTs in the enumeration conditions. This resulting variable is below referred to as the enumeration time as it is not confounded by RT differences in pronouncing different number words. Enumeration times (correct responses within 150–3000 ms from stimulus onset) were analyzed using linear mixed-effects models. The advantage of mixed-effects models is that it enables the examination of individual differences (e.g., in the subitizing slope) in addition to group-level effects. The model was fit (maximized log-likelihood) using the nlme package (Pinhero, Bates, DebRoy, Sarkar, & R Core Team, 2014; for an introduction to mixed-effects models, see Bayeen, Davidson, & Bates, 2008). In the mixed-effect models, the number of items (1–6), background type (full scene vs. uniform background), and their interaction were added as fixed-effects factors. To model the breakpoint in enumeration performance, the data was fitted with two linear segments, the first segment for the subitizing range and the second segment for higher numbers. Participants were defined as random effects so that each participant's data were fitted with individual intercepts and slopes, separately for each linear segment. The model is presented formally in detail in the Appendix. The use of participant-wise random effects was useful due to the variation between participants' enumeration performance. To find the upper bound of the subitizing range, we fitted three alternative models where the length of the first segment ranged numbers 1–2, 1–3, or 1–4. In order to test whether the enumeration data contains a breakpoint in performance, also a simple linear model that only consisted of one segment was estimated. Model fits were evaluated using the Akaike Information Criterion,

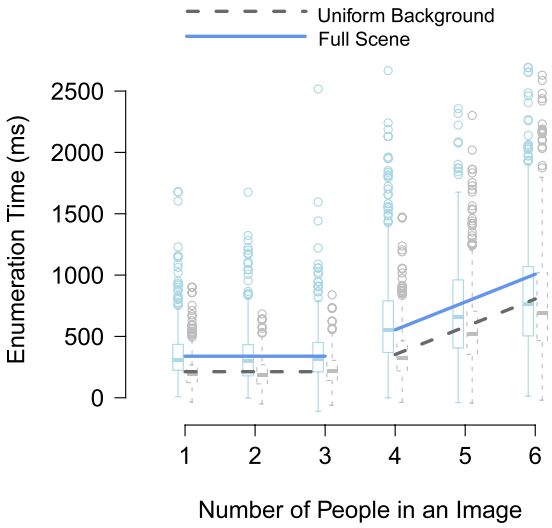


Figure 2. Enumeration times as a function of number and background condition. The boxplots show the observed data and the lines represent the fitted linear mixed-effects model. Blue color (solid line) denotes the full scene condition and gray (dashed line) the uniform background condition.

which weighs the goodness of fit by the complexity of the model. Different models were compared using likelihood-ratio tests. Due to right-skewed RT distributions the mixed-model analysis was also performed on log-transformed data, but this did not significantly change the results. We report the results of the tests performed on untransformed data to make the interpretation of the results easier.

Enumeration accuracies were analyzed using mixed-effects logit models (Jaeger, 2008) with binomial probability distributions and logit link function, using the R package lme4 (Bates, Maechler, Bolker, & Walker, 2014). The fixed- and random-effects terms of the full model correspond to the linear mixed-effects model presented above (and in the Appendix). As with RT analyses, two-segment models were used to model the breakpoint in enumeration accuracies.

Datasets and the analysis script can be downloaded at the Open Science Framework (https://osf.io/rtdfe/).

# Results

## Enumeration times

As shown in Figure 2, enumeration time data revealed the characteristic subitizing-counting breakpoint in both full scene and uniform background conditions. A two-segment model where the length of the first segment spanned numbers 1–3 yielded the best

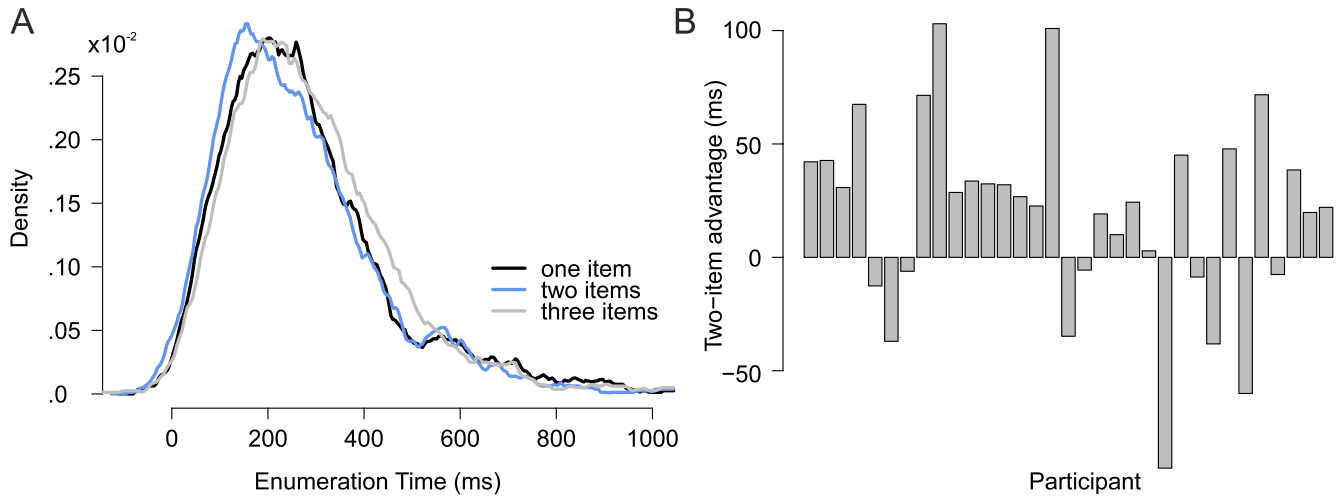


Figure 3. (A) Kernel density estimates (calculated using function density in *R*; bandwidth = 25 ms) of enumeration time distributions for numbers 1–3 (averaged across full scene and uniform background conditions). A small proportion of enumeration times are negative because enumeration time was calculated by subtracting the number-naming RT from single-trial enumeration RTs (note that the shape of the distribution is not affected by this transformation, only its position changes). (B) Two-item advantage (mean two-item enumeration time subtracted from mean one-item enumeration time) plotted as a function of participant.

fit (comparison against corresponding one-segment model: L.Ratio = 819.5, $p < 0.0001$). Below, the first segment is referred to as the subitizing range and the second segment as the counting range. Enumeration times remained constant in the subitizing range ($\beta = 8.13$, $t = 1.36$, $p = 0.17$), and there was no interaction between number (1–3) and condition ($\beta = -13.37$, $t = -1.51$, $p = 0.13$). Next, the full model was simplified by removing nonsignificant fixed-effect regressors and unnecessary random-effect terms. This pruned model is shown in Figure 2 ($df = 4191$). The average speed of enumerating a single item in the uniform background condition (i.e., intercept of the first segment) was 212 ms (CI of the intercept = [177.5, 247.3], $t = 11.93$, $p < 0.0001$). This effect was adjusted by the random-effect term to take into account participant-wise variation in enumeration speed ($SD = 98.05$ ms, CI of the $SD$ = [75.31, 127.64] ms). As the number of objects did not statistically significantly modulate enumeration times in the subitizing range, the intercept reflects the estimated enumeration time for each number in the subitizing range. This effect was consistent across participants, as including participant-wise random variation in the slope of the first segment into the model did not increase the fit of the model (and was therefore excluded from the model). Outside the subitizing range, enumeration of four items on uniform background took on average 139.96 ms (intercept of the second segment; CI = [109.9, 169.3]) longer than enumerating items on uniform background in the subitizing range ($t = 9.15$, $p < 0.0001$; random-effect variation: $SD = 70.22$ ms, CI = [49.71, 99.18] ms). Enumeration times also increased as number increased in the counting range ($\beta = 226.56$, CI = [180.6, 272.5], $t = 9.67$, $p < 0.0001$). In contrast to the flat slope in subitizing range, there was substantial variation between participants in the counting range intercept ($SD = 70.2$, CI = [49.6, 99.4] ms) and slope ($SD = 129.84$, CI = [99.01, 170.16] ms). When the images contained the natural background, enumeration times were delayed by 126.19 ms on average (CI = [89.8, 162.6], $t = 6.80$, $p < 0.0001$) in both subitizing and counting ranges (participant-wise random variation: $SD = 98.19$ ms, CI = [74.34, 128.65] ms). An interaction between background condition and the intercept of the second segment showed that when the images contained a natural background, enumeration times in the counting range increased an additional 75 ms, when compared to the subitizing range (CI = [50.7, 99.9], $t = 6.00$, $p < 0.0001$).

The above results suggest that enumeration times remain constant in the subitizing range. A closer look at the enumeration times in the subitizing range (Figure 3A) showed that two items were enumerated faster (269.59 ms, $SE = 21.36$) than one item (288.75 ms, $SE = 19.79$; mean difference = 19.2 ms, $t = 2.6$, $df = 32$, $p = 0.014$; due to the lack of subitizing range × background condition interaction, the full scene and uniform background conditions were pooled for this analysis). As shown in Figure 3B, this effect was observed in 23 out of 33 participants. A similar effect has been previously reported and called the two-item advantage (Railo, 2014). To further examine this effect, enumeration times in the one- and two-item conditions were estimated by linear mixed-effects models. In order to control for possible confounding factors, the RT of the preceding trial was brought into the model as a

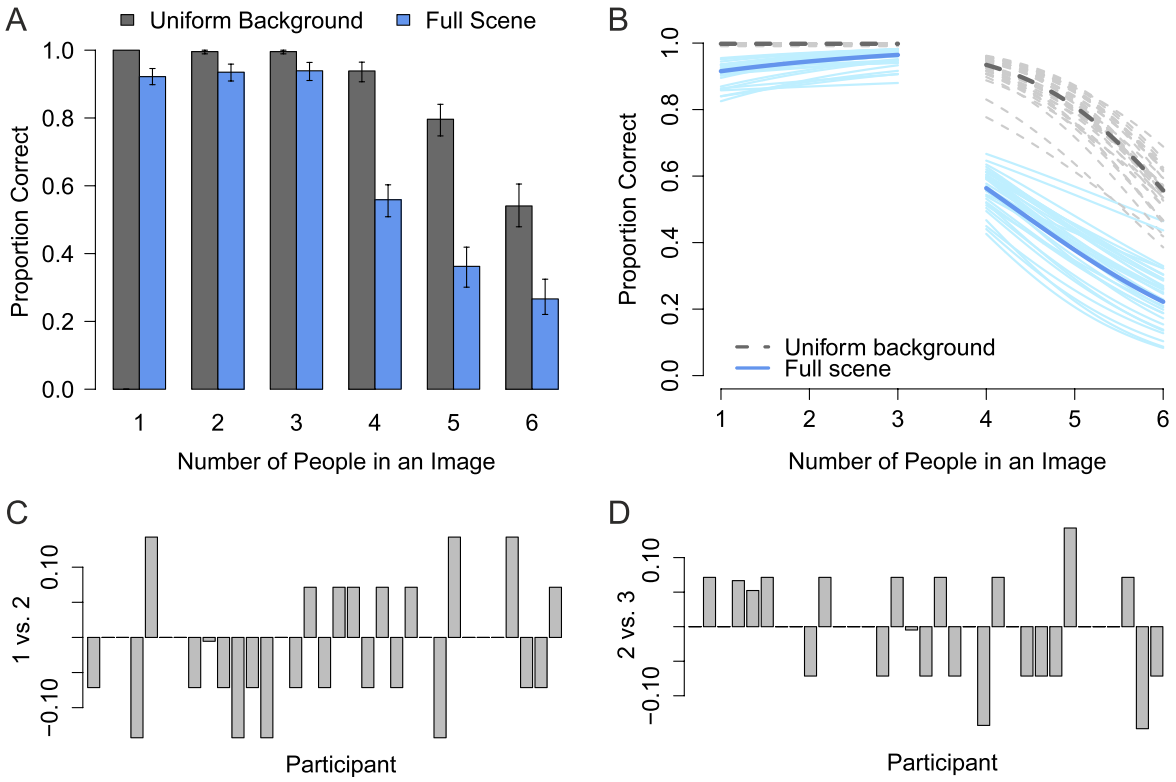


Figure 4. (A) Mean enumeration accuracy as a function of number and background condition. Error bars represent the 95% CIs (based on 1,000 bootstrap samples). No error bar is shown for the one-item uniform background condition because enumeration was always accurate. (B) Results of the mixed logit model. The narrow, light-colored lines depict single participants' results, and the thicker, more distinct lines show the group results. (C) Difference in enumeration accuracy in the full scene one versus two-item conditions as a function of participant. Positive values denote that enumeration accuracy was higher in the one-item condition. (D) Difference in enumeration accuracy in the full scene two- versus three-item conditions as a function of participant. Positive values denote that enumeration accuracy was higher in the two-item condition. Similar data is not presented for the uniform background condition because there were very few incorrect answers.

predictor, in addition to number of items (1 vs. 2) and condition (number × condition interaction did not reach statistical significance [$p = 0.49$], and was thus left out of the model). Previously, the RTs in the preceding trials have been shown to influence RTs (Maljkovic & Nakayama, 1994). It is important to control for this in the present context as it could be argued that the preceding trial influences the participants' expectations about the number of stimuli presented next. As over 80% of the trials contained at least two items in the present study, preceding trials could influence, or even produce the two-stream advantage. The results ($df = 1751$) showed that across participants enumerating two-items was 16.2 ms faster than enumerating one item (CI = [29.4, 3.01], $t = -2.40$, $p = 0.016$), even when the influence of the previous trial RT was controlled for ($\beta = .085$, CI = [0.06–0.10], $t = 8.6$, $p < 0.0001$). Participant-wise random intercepts were included in the model ($SD = 100.1$ ms, CI = [77.8, 128.8] ms) but adding random slopes did not improve the fit of the model ($p = 0.2$). As before, the effect of background

| | Fixed effects | | | Random effects |
|---|---|---|---|---|
| | OR | 95% CI | *Z* value | *SD* |
| Intercept (subitizing) | 473.2 | [232.0, 965.0] | 16.9 | 2.6 |
| Bgr | 0.02 | [0.01, 0.04] | −9.5 | 2.4 |
| Nbr (subitizing) × Bgr | 1.6 | [1.3, 1.9] | 4.9 | 1.2 |
| Intercept (counting) | 0.03 | [0.01, 0.05] | −15.6 | 1.9 |
| Nbr (counting) | 0.3 | [0.02, 0.06] | −13.2 | 1.3 |

Table 2. Results of the mixed-effects logit model ($df = 5416$). *Notes*: OR = odds ratio; Bgr = background condition; Nbr = number.

condition was statistically significant ($\beta = 139.5$, CI = [126.0–153.0], $t = 20.3$, $p < 0.0001$).

### Enumeration accuracies

As shown in Figure 4A, enumeration was accurate within the subitizing range, and decreased in the counting range. As with enumeration times, a two-segment model where the first segment ranged numbers 1–3 fit the data best (comparison against corresponding one-segment model: L.Ratio = 103.6, $p < 0.0001$). The best fitting two-segment model was pruned by removing nonsignificant terms. The results are presented in Table 2, and the model is visualized in Figure 4B. All effects are highly statistically significant ($p < 0.0001$). The model revealed an interaction between number and background condition in the subitizing range: Whereas accuracy remained almost perfect in the uniform background condition, accuracy slightly increased as number increased in the full scene condition. In the counting range, enumeration accuracies decreased in general relative to the subitizing range (intercept of the counting range), and the probability of correct response decreased as number increased (main effect of number in counting range). Enumeration was more error prone in the full background than in the uniform background condition.

Figure 4C and D present differences in enumeration accuracy between numbers 1 versus 2 and 2 versus 3 in the full-scene condition (observed data). The participants who displayed a two-item advantage in RTs in the full-scene condition also showed a marginally statistically significant tendency to enumerate two items more accurately than one item in the whole scene condition (Pearson's $r = -0.34$, $p = 0.055$).

## Discussion

We found that enumeration of items presented in their natural backgrounds follows a similar two-segment function as enumeration of items in uniform background: Enumeration performance remained approximately constant for up to three items (subitizing range), while the enumeration of larger numbers was associated with increasing costs (counting range). Thus, the results show that up to three visual items can be enumerated through subitizing when the items are embedded in real-world contexts. Importantly, the subitizing range and slope was similar in the natural scene condition when compared to enumerating the same objects on a uniform background. However, when the items were embedded in background, enumeration times were delayed by ~125 ms in the subitizing range. This delay is likely related to parsing out the humans from the background. The constant delay suggests that small numbers of humans were segregated from the background and detected without additional processing costs. If detecting the targets from the background would have required serial attention, the delay should have increased as the number of objects increased. Similarly, using naturalistic stimuli, Wolfe et al. (2002) showed that increasing background complexity produces a strong additive change to visual search RTs. Previous findings using artificial stimuli also show that figure–ground segmentation can take place with minimum attention (Kimchi & Peterson, 2008). However, our results do not imply that figure–ground segregation takes place without any attention. In the counting range, the presence of a background increased processing times 75 ms more than in the subitizing range. This suggests that figure–ground segmentation is dependent on similar capacity limitations as enumeration.

Because the low-level information of the stimuli were matched in the present study, it is very unlikely that the participants could have determined the number of humans solely based on low-level visual information. However, the detection of target objects could be mediated by feature detectors that are tuned to learned features of intermediate complexity such as faces (Evans & Treisman, 2005; Ullman, Vidal-Naquet, & Sali, 2002). Future research should examine whether the present results generalize to situations where participants enumerate other objects than humans. Fei-Fei et al. (2007), whose results suggest that humans show a preference for perceiving animate objects (such as humans), hypothesize that "there might also be efficient computational mechanisms for the visual system to process this information rapidly and accurately" (p. 24). Thus, the category of items that are enumerated might influence the results. Finally, it should also be noted that the capacity to detect (and enumerate) objects also depends on how well the stimuli stand out from the background (Wolfe et al., 2002).

Against serial and parallel multifocal models of multiple object processing (Cavanagh & Alvarez, 2005; Huang & Pashler, 2007; Oksama & Hyönä, 2008), the present results show that participants are, on average, fastest in enumerating two objects—enumerating a single object required additional processing time. Similarly, enumeration became less error prone as number increased in the subitizing range, suggesting that at group level, the participants were tuned to process multiple items. This effect was only observed in the full-scene condition, perhaps because enumeration accuracy was at ceiling in the uniform background condition. Both effects showed individual variation. The correlation between the two-item advantage and a

corresponding difference in enumeration accuracies suggests that the two effects may be related, and that they do not reflect speed–accuracy trade-off.

The two-item advantage may partly be explained by the assumption that two items can be independently represented in the two hemispheres (Alvarez & Cavanagh, 2005; Alvarez, Gill, & Cavanagh, 2012) as the two-item advantage is more pronounced when using bilateral than unilateral stimuli (Railo, 2014), and in the present study the objects were typically uniformly distributed in both hemifields. The two-item advantage and superior accuracy for enumerating multiple items in the subitizing range may also be specifically related to the way objects are treated by attentional mechanisms in subitizing. Vuilleumier and Rafal (1999) showed that when hemispatial neglect patients localized or enumerated one versus two visual targets, contralateral extinction was greatly reduced in the enumeration condition. The authors hypothesized "that enumeration allowed linkage rather than competition between bilateral stimuli," enabling the patients to see one group of two stimuli rather than two separate stimuli (Vuilleumier & Rafal, 1999, p. 784). Subitizing (Mandler & Shebo, 1982) and object representations in general (Alvarez, 2011) have been suggested to represent multiple objects as one integrated ensemble to facilitate processing. This suggests that the attentional requirements in subitizing come from constructing an integrated conscious percept of individual objects. Because the enumeration task typically requires participants to process multiple items, top-down attentional predictions may expect multiple items to be presented (e.g., Panichello, Cheung, & Bar, 2012). These top-down predictions may facilitate the processing of multiple item displays, but impede the processing of singular items due to a mismatch between top-down predictions and bottom-up visual information (Railo, 2014).

The observed advantage concerning processing multiple items is consistent with the proposal that the representation of small numbers of objects is hardwired into humans (and animals; Hauser & Carey, 2003), and comprises one core system on which the development of numerical cognition is founded on (Spelke, 2011; Spelke & Kinzler, 2007). One may also speculate that the ability to simultaneously and preconceptually individuate two items may serve as the basis for understanding spatial relations (e.g., whether an object is on top of another object; Pylyshyn, 2001), and thereby underlie inherent spatial intuitions (Dehaene, Izard, Pica, & Spelke, 2006). The ability to simultaneously detect multiple objects could also mediate the ability of humans to grasp what is happening in complex visual scenes from a single glance. For example, Hafri, Papfragou, and Trueswell (2013) showed that participants can recognize what is happening in briefly presented natural images depicting various interactions between two humans. However, further studies are essential to verify and interpret the two-item advantage and the increase in enumeration accuracy in the subitizing range. A central open question is whether the observed results generalize to other tasks than enumeration. Is it observed in tasks that require participants to make decisions about integrated collections of items rather than individual items, is it produced by top-down expectations (Railo, 2014), or could it reflect a more permanent organization of mental representations (Mandler, 2013)?

In summary, the main finding of the present study is that multiple objects can be segregated from the ground and up to three objects can be simultaneously bound into an integrated percept when viewing naturalistic stimuli. Moreover, the visual system appears to be tuned to processing multiple items, which may underlie spatial and numerical cognition, and be beneficial in various real-world situations that often require perceiving multiple objects.

*Keywords: subitizing, scene perception, natural scene categorization, number sense, visual attention, figure–ground segmentation*

## Acknowledgments

We thank anonymous reviewers for their helpful comments. HR was funded by Turku Institute of Advanced Studies.

Commercial relationships: none.
Corresponding author: Henry Railo.
Email: hmrail@utu.fi.
Address: Department of Psychology, University of Turku, Turku, Finland.

## References

Alvarez, G. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131.

Alvarez, G. A., & Cavanagh, P. (2005). Independent resources of attentional tracking in the left and right visual hemifields. *Psychological Science*, *16*(8), 637–643.

Alvarez, G. A., Gill, J., & Cavanagh, P. (2012). Anatomical constraints on attention: Hemifield independence is a signature of multifocal spatial selection. *Journal of Vision*, *12*(5):9, 1–20, doi:10.1167/12.5.9. [PubMed] [Article]

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (R package Version 1.1-7). Retrieved from http://CRAN.R-project.org/package=lme4

Bayeen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, *9*(7), 349–354.

Cohen, M. A., Alvarez, G. A., & Nakayama, K. (2011). Natural-scene perception requires attention. *Psychological Science*, *22*(9), 1165–1172.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an amazonian indigene group. *Science*, *311*, 381–384.

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1476–1492.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1):10, 1–29, doi:10.1167/7.1.10. [PubMed] [Article]

Fei-Fei, L., VanRullen, R. Koch, C., & Perona. P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, *12*(6), 893–924.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307–314.

Hafri, A., Papfragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, *142*(3), 880–905.

Hauser, M., & Carey, S. (2003). Spontaneous representations of small numbers of objects by rhesus macaques: Examinations. *Cognitive Psychology*, *47*, 367–401.

Huang, L., & Pashler, H. (2007). A Boolean map theory of attention. *Psychological Review*, *114*(3), 599–631.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jevons, W. S. (1871). The power of numerical discrimination. *Nature*, *3*, 281–282.

Kaufman, E., Lord, M. Reese, T., & Volkmann, J. (1949). The discrimination of visual number. *American Journal of Psychology*, *62*, 498–525.

Kimchi, R., & Peterson, M. A. (2008). Figure-ground segmentation can occur without attention. *Psychological Science*, *19*(7), 660–668.

Li, F. F., Van Rullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, B*, *99*, 9596–9601.

Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory and Cognition*, *22*, 657–672.

Mandler, G. (2013). The limit of mental structures. *The Journal of General Psychology*, *140*(4), 243–250.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, *111*(1), 1–22.

Oksama, L., & Hyönä, J. (2008). Dynamic serial binding of identity and location information: A serial model of multiple object tracking. *Cognitive Psychology*, *56*(4), 237–283.

Panichello, M. F., Cheung, O. S., & Bar, M. (2012). Predictive feedback and conscious experience. *Frontiers in Psychology, 3*, 620, doi:10.3389/fpsyg.2012.00620.

Piazza, M., Fumarola, A., Chinello, A., & Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition*, *121*(1), 147–153.

Pinhero, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2014). nlme: Linear and nonlinear mixed effects models. Retrieved from http://CRAN.R-project.org/package=nlme

Pylyshyn, Z. W. (2001). Visual indexes, preconscious objects, and situated vision. *Cognition*, *80*(1–2), 127–158.

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

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org/

Railo, H. (2014). Bilateral and two-item advantage in subitizing. *Vision Research*, *103*, 41–48.

Railo, H., Koivisto, M., Revonsuo, A., & Hannula, M. M. (2008). The role of attention in subitizing. *Cognition*, *107*, 82–104.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*(6), 607–614.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, *5*(7), 629–630.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004a). How parallel is visual processing in the ventral visual pathway? *Trends in Cognitive Sciences*, *8*(8), 363–370.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004b). Processing of one, two or four natural scenes in humans: The limits of parallelism. *Vision Research*, *44*, 877–894.

Spelke, E. (2011). Natural number and natural geometry. In S. Dehaene & E. Brannon (Eds.), *Space, time and number in the brain* (pp. 287–317). London: Academic Press.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Psychology*, *10*(1), 89–96.

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

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*(1), 80–102.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682–688.

VanRullen, R., & Koch, C. (2003). Competition and selection during visual processing of natural scenes and objects. *Journal of Vision*, *3*(1):8, 75–85, doi:10.1167/3.1.8. [PubMed] [Article]

VanRullen, R., Reddy, L., & Fei-Fei, L. (2005). Binding is a local problem for natural objects and scenes. *Vision Research*, *45*, 31333–3144.

VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual task reveal two distinct attentional resources. *Journal of Cognitive Neuroscience*, *16*(1), 4–14.

Vetter, P., Butterworth, B., & Bahrami, B. (2008). Modulating attentional load affects numerosity estimation: Evidence against a pre-attentive subitizing mechanism. *PLoS ONE*, *3*, e3269.

Vuilleumier, P., & Rafal, R. (1999). Both means more than "two": Localizing and counting in patients with visuospatial neglect. *Nature Neuroscience*, *2*(9), 783–784.

Walker, S., Stafford, P., & Davis. G. (2008). Ultrarapid categorization requires visual attention: Scenes with multiple foreground objects. *Journal of Vision*, *8*(4):21, 1–12, doi:10.1167/8.4.21. [PubMed] [Article]

Watson, D. G., Maylor, E. A., Allen, G. E., & Bruce, L. A. (2007). Early visual tagging: Effects of target-distractor similarity and old age on search, subitization, and counting. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(3), 549–569.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Grosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, *42*(3), 671–684.

Wolfe, J. M., Oliva, A., Horowitz, T. S., Butcher, S. J., & Bombas, A. (2002). Segmentation of objects from background in visual search tasks. *Vision Research*, *42*, 2985–3004.

## Appendix

Data was analyzed using a mixed-effects regression models that consisted of two linear segments. The largest form of the model that we consider is:

$$\begin{aligned} Y_{ij} = {} & \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} \\ & + \beta_5 x_{1ij} x_{4ij} + \beta_6 x_{2ij} x_{4ij} + \beta_7 x_{3ij} x_{4ij} + b_{i0} \\ & + b_{i1} x_{1ij} + b_{i2} x_{2ij} + b_{i3} x_{3ij} + b_{i4} x_{4ij} \\ & + b_{i5} x_{1ij} x_{4ij} + b_{i6} x_{2ij} x_{4ij} + b_{i7} x_{3ij} x_{4ij} + \epsilon_{ij} \end{aligned} \quad (1)$$

After removing unnecessary fixed-effect and random-effect terms the linear mixed-effects model used to estimate RTs is simplified to:

$$\begin{aligned} Y_{ij} = {} & \beta_0 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_6 x_{2ij} x_{4ij} + b_{i0} \\ & + b_{i2} x_{2ij} + b_{i3} x_{3ij} + b_{i4} x_{4ij} + \epsilon_{ij} \end{aligned} \quad (2)$$

Mixed-effects logit models were used to analyze enumeration accuracies. Below, $Y_{ij}$ refers to the probability of correct response. The simplified form of the mixed-effects logit is:

$$\begin{aligned} logit(Y_{ij}) = {} & \beta_0 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{1ij} x_{4ij} \\ & + b_{i0} + b_{i2} x_{2ij} + b_{i3} x_{3ij} + b_{i4} x_{4ij} \\ & + b_{i5} x_{1ij} x_{4ij} + \epsilon_{ij} \end{aligned} \quad (3)$$

where

– $Y_{ij}$: the $j$th observation of $i$th participant
– $x_{1ij}$: the number of people in an image –1
– $x_{2ij}$: dichotomous variable denoting that the number of people in an image is equal to or larger than $T$, i.e., the starting point of the second segment

- $-x_{3ij}$: the number of people in an image in addition to $T$
- $-x_{4ij}$: a dichotomous variable denoting that the image contains a background
- $-\beta_0 \ldots \beta_7$: the fixed-effects coefficients:
  - $-\beta_0$: the mean of the dependent variable in an image with one person with a uniform background
  - $-\beta_1$: change when adding one more person in an image with a uniform background
  - $-\beta_2$: change when there is at least $T$ people in an image with a uniform background
  - $-\beta_3$: change when adding one more person in an image with a uniform background when there is at least $T$ people in the image
  - $-\beta_4$: change when the image contains full back ground
  - $-\beta_5$: change when adding one more person in full background image
  - $-\beta_6$: change in the mean of the dependent variable when there is at least $T$ people in an image with a full background
  - $-\beta_7$: change when adding one more person in an image with a full background when there is at least $T$ people in the image
- $-b_{i0} \ldots b_{i7}$: random-effect coefficient for $i$th participant
- $-\epsilon_{ij}$ is the error for observation $j$ of participant $i$