Cognition 228 (2022) 105227

Contents lists available at ScienceDirect

Cognition

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

Capacity limits in face detection

Rana Qarooni [a], Jonathan Prunty [b], Markus Bindemann [b], Rob Jenkins [a,*]

[a] Department of Psychology, University of York, UK
[b] School of Psychology, University of Kent, Canterbury, UK

ARTICLE INFO

Keywords:
Face detection
Face perception
Capacity limits
Visual attention

ABSTRACT

Face detection is a prerequisite for further face processing, such as extracting identity or semantic information. Those later processes appear to be subject to strict capacity limits, but the location of the bottleneck is unclear. In particular, it is not known whether the bottleneck occurs before or after face detection. Here we present a novel test of capacity limits in face detection. Across four behavioural experiments, we assessed detection of multiple faces via observers' ability to differentiate between two types of display. *Fixed* displays comprised items of the same type (all faces or all non-faces). *Mixed* displays combined faces and non-faces. Critically, a 'fixed' response requires all items to be processed. We found that additional faces could be detected with no cost to efficiency, and that this capacity-free performance was contingent on visual context. The observed pattern was not specific to faces, but detection was more efficient for faces overall. Our findings suggest that strict capacity limits in face perception occur after the detection step.

1. Introduction

Studies of face perception often emphasise the wealth of social information that we derive from faces. However, access to this information is gated by the prior step of face detection, in which the visual system registers the presence of a face. Despite its gatekeeper role, face detection has received little research attention compared to later stages of face perception (e.g. identification; social inferences). Most of the research on face detection concerns algorithm development in computer vision (see Hjelmås & Low, 2001; Kumar, Kaur, & Kumar, 2018 for reviews). As such, the cognitive process of face detection is not well understood.

We follow previous researchers in assuming that face detection involves matching a region of the visual field to a stored face template (Lewis & Ellis, 2003; Robertson, Jenkins, & Burton, 2017; Tsao & Livingstone, 2008). The few psychological studies that have addressed this process have tended to focus on qualitative aspects of the putative template, such as sensitivity to the colour, outline, or spatial layout of the face (Amso, Haas, & Markant, 2014; Bindemann & Burton, 2009; Crouzet, Kirchner, & Thorpe, 2010; Pongakkasira & Bindemann, 2015; Purcell & Stewart, 1988; Purcell & Stewart, 1986; Simpson, Maylott, Leonard, Lazo, & Jakobsen, 2019; Stein, Sterzer, & Peelen, 2012). Less still is known about quantitative aspects of face detection, such as whether multiple faces can be detected at once. In some ways this is a puzzling omission, as quantitative aspects of later face perception processes (e.g. gaze perception, identification, semantic association) have been studied in some detail (Bindemann & Burton, 2009; Bindemann, Burton, Hooge, Jenkins, & De Haan, 2005; Bindemann, Jenkins, & Burton, 2007; Jenkins, Lavie, & Driver, 2003).

Several of those studies have recruited the notion of *capacity limits*—the basic observation that not all the available sensory information can be processed at once (Bruckmaier, Tachtsidis, Phan, & Lavie, 2020; Lavie & De Fockert, 2003; Norman & Bobrow, 1975). The further claim is that face processing may be subject to its own, face-specific capacity limits (Jenkins et al., 2003). Surprisingly, this limit may be as low as a single face, such that face processing proceeds one face at a time.

The evidence leading to this claim comes from a range of behavioural experiments. For example, patterns of response competition effects (Bindemann, Burton, et al., 2005; Jenkins et al., 2003; Thoma & Lavie, 2013) and repetition priming effects (Bindemann et al., 2007) indicate that processing one face selectively blocks processing of another face. This finding applies not only to later, cognitively deep processes involving extraction of personal identity or semantic information (Bindemann et al., 2007; Bindemann, Burton, & Jenkins, 2005; Jenkins et al., 2003; Thoma & Lavie, 2013), but also to earlier, cognitively shallow processes such as classifying sex (male/female, Bindemann, Burton, & Jenkins, 2005) or gaze direction (left/right, Bindemann & Burton, 2009).

* Corresponding author at: Department of Psychology, University of York, York YO10 5DD, UK.
E-mail address: rob.jenkins@york.ac.uk (R. Jenkins).

https://doi.org/10.1016/j.cognition.2022.105227
Received 25 January 2022; Received in revised form 13 July 2022; Accepted 14 July 2022
Available online 21 July 2022

Together, these findings point to a bottleneck early in face processing (i.e. upstream of sex or gaze perception) that constrains face processing downstream of the bottleneck. One possibility is that face detection itself is the bottleneck. This possibility implies strict capacity limits at the detection stage, such that only one face at a time can be acquired from the visual environment (though still allowing rapid serial acquisition). Alternatively, detection itself could be capacity free. This possibility implies that multiple faces can be acquired in parallel, they just cannot be processed in parallel. On this view, the bottleneck occurs when extracting information from faces.

Can we know that more than one face is present? ERP experiments offer some evidence on this point. The amplitude of the N170, a face-selective ERP marker, has been found to increase when multiple faces are presented (Puce et al., 2013). However, as the authors acknowledge, their task of reporting the number of stimuli (1–3) did not require participants to know whether or not the stimuli were faces. Registering the presence of any stimuli could produce the same results.

Several behavioural studies have addressed the related question of whether a target face 'pops out' from surrounding distractors in a visual search task (Brown, Huey, & Findlay, 1997; Kuehn & Jolicoeur, 1994; Lewis & Edmonds, 2005; Nothdurft, 1993; Treisman & Gelade, 1980). However, these studies have led to conflicting results. For instance, Nothdurft (1993) found that search times increased with the number of distractors (set size), suggesting serial processing. In contrast, Lewis and Edmonds (2005) found equivalently low search times regardless of set size, suggesting parallel processing. We return to this discrepancy in the General Discussion section.

Although visual search can be informative, there are several reasons why it may be unsuitable for probing capacity limits in face detection. First, the task imposes a distinction between target and distractor stimuli. This distinction gives special status to the target category, potentially affecting attentional set (Bindemann et al., 2007; Wolfe & Horowitz, 2004). Second, visual search entails active scanning for the target (Jenkins et al., 2003), whereas everyday face detection often occurs incidentally during passive viewing. Third, and most importantly for the current study, visual search does not lend itself to testing detection of *multiple* faces. As the participant's task is to indicate the presence or absence of a target, search can be terminated when a single target is found, even if other targets are present.

Previous researchers have distinguished two components of face detection that are sometimes conflated — localising and categorising (Bindemann & Lewis, 2013). The localising component involves searching for a target under spatial uncertainty. Categorisation involves establishing whether or not a stimulus is a face. As we are primarily interested in template matching, our focus here was the categorisation aspect of detection. To probe capacity limits, we sought to assess the cognitive cost of increasing the number of items, while eliminating any cognitive costs associated with localising those items.

To this end, we devised a new task in which all items have equal status. Participants saw face and non-face items in 'fixed' displays (all one stimulus type) or 'mixed' displays (a combination of both types). To reduce the need for visual scanning, these items appeared at predefined locations surrounding central fixation at low eccentricities (that is, with spatial certainty for addressable coordinates; Garner, Bowman, & Raymond, 2021). For each display, the participants' task was to indicate whether the items were fixed or mixed. Critically, this task involves assimilating multiple faces. In particular, correct 'fixed' responses require each item to match (or not match) a face detection template before a response is made. We take reportability via this fixed/mixed judgement as our detection criterion. Manipulating the type and number of items in the display allows us to estimate per-item detection costs separately for each stimulus type. We take positive cost per item to indicate that detection is capacity limited. Conversely, we take zero cost per item to indicate that detection is capacity free.

2. Experiment 1

We began by comparing detection efficiency for faces and non-faces presented in 'fixed' or 'mixed' displays of set size two or three. Non-faces in this experiment were scrambled (phase-shifted) faces that matched the low-level visual energies of the intact face stimuli (Jenkins et al., 2003). For set size two, *Fixed* conditions contained either two faces (FF) or two non-faces (NN), while *Mixed* conditions contained one stimulus of each type (FN or NF, differentiated by spatial layout). Set size three conditions were constructed by adding an extra face or non-face to the display. Thus, *Fixed* conditions contained either three faces (FFF) or three non-faces (NNN), while *Mixed* conditions combined both types of stimulus (FNN and NFF). Participants were asked to decide as quickly and accurately as possible whether each display was *Fixed* (all the same type of stimulus), or *Mixed* (a combination of both stimulus types). Comparing set size two and three allowed us to estimate the effect of an extra display item on these determinations. We expected that capacity-limited face detection should enforce serial processing, resulting in greater efficiency for set size two than for set size three. On the other hand, capacity-free face detection should allow parallel processing, resulting in equivalent efficiency for set size two and for set size three.

2.1. Methods

2.1.1. Participants

Seventy-seven participants were recruited through Prolific recruitment service (www.prolific.co) and completed the experiment in exchange for a small payment. Seventeen participants were excluded due to failed attention checks (as described below; exclusion criteria: 2 or more failed checks within the same block, or 3 across the whole experiment) or slow responses (>2.5 SD from the group mean). The final sample ($N = 60$) comprised 22 females and 38 males (age range 18–73; M = 27.95, SD = 11.09).

2.1.2. Design and stimuli

Stimuli were generated using a local face bank of 288 faces. The local bank consisted of AI generated faces (Karras and Nvidia, 2018) supplemented with real faces from the MR2 face bank (Strohminger et al., 2016) and other online sources. This local face bank contained an equal distribution of age (younger adults and older adults), sex (male and female), and ethnicity (Asian, Black, and Caucasian; see Prunty et al. (under review) for details of demographic categorisation). Each image was cropped to a 380-pixel wide × 570-pixel high rectangle to create the face items. To create the non-face items, each face was submitted to Fourier phase transformation that randomly scrambled the phase of component spatial frequencies while maintaining overall brightness, contrast, and orientation (Honey, Kirchner, & VanRullen, 2008). Fig. 1 shows examples of this manipulation.

To construct Set Size *Two* displays, we used 58 randomly selected faces from the local face bank together with their scrambled face counterparts. For Set Size *Three*, we used 87 faces and scrambled faces. No item was repeated within a display or within a condition. The selected items were randomly allocated to three predetermined locations that formed an upright equilateral triangle around central fixation (nearest contours ~95 pixels apart).

Catch trials for each condition were created for use as attention checks. These catch trials were constructed in a similar manner and used the same spatial layout. Intact faces were substituted with a white vertical rectangle containing a smaller black rectangle in the centre. Scrambled faces were replaced with a phase-scrambled equivalent of these rectangles.

The within-subjects factors of Display Type (*Fixed*, *Mixed*) and Set Size (*Two*, *Three*) were manipulated in a fully counterbalanced 2 × 2 factorial design, resulting in the experimental conditions summarised in Fig. 1a.

The experiment was created and hosted online at Gorilla Experiment

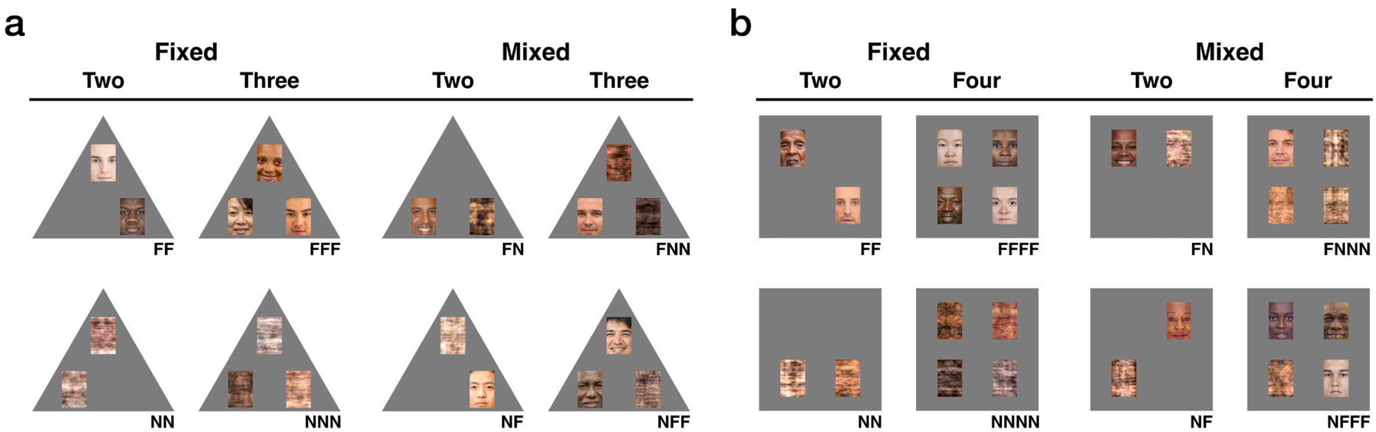

Fig. 1. Example displays for each condition of (a) Experiment 1 and (b) Experiment 2. Fixed displays contained one type of stimulus (all faces or all non-faces). Mixed displays contained both types. Numbers refer to set sizes. F denotes face, N denotes non-face (scrambled faces in Experiments 1 & 2). Triangle and square segments are for visualisation only. In the actual experiments, the grey background filled the whole screen.

Builder (gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). Participants could access the experiment on any desktop or laptop computer, precluding exact control over screen size. Mobile devices and tablets were excluded.

2.1.3. Procedure

Participants were asked to indicate as quickly and accurately as possible whether items in a display were *Fixed* and of the same type (i.e., all face or all non-face), or *Mixed* and a combination of both types (i.e., faces and non-faces together). Each trial started with a fixation cross for 250 ms followed by stimulus displays presented until response. The experiment began with a practice block of 16 trials consisting of two different trials per condition in random order. This was followed by 3 experimental blocks, each consisting of 72 experimental trials (9 trials per conditions) plus 4 catch trials in a random order. Participants were given the opportunity to take short breaks between the blocks. The entire experiment took approximately 10 min to complete.

2.2. Results and discussion

Overall accuracy for the Fixed/Mixed judgements was 95%, confirming that participants could distinguish between face and non-face stimuli. Trials with reaction times below 150 ms or above 3000 ms were excluded from analysis (0.57% of all trials). For concision, we combined accuracy and reaction time data to form Linear Integrated Speed-Accuracy Scores (LISAS) which summarise performance in a single efficiency metric while accounting for speed-accuracy trade-offs in responses (Vandierendonck, 2017, 2018, 2021). Separate analyses of accuracy and reaction time measures for each experiment are reported in Supplementary Materials and support the same conclusions.

Fig. 2a summarises Linear Integrated Speed-Accuracy Scores for each condition in Experiment 1. LISAS data were submitted to a two-way ANOVA with the repeated measures factors of Set Size (*Two*, *Three*) and Display Type (*Fixed*, *Mixed*). This analysis revealed a significant main effect of Set Size, with more efficient detection for *Two* items (M = 768 ms, SE = 9 ms) than for *Three* items (M = 785 ms, SE = 10 ms) overall [$F(1, 59) = 9.16$, $p = .004$, $\eta^2 = 0.31$], and a significant main effect of Display Type, with more efficient detection for *Fixed* displays (M = 756 ms, SE = 10 ms) than for *Mixed* displays (M = 797 ms, SE = 9 ms) overall [$F(3, 177) = 13.08$, $p < .001$, $\eta^2 = 0.18$]. There was also a significant interaction between Set Size and Display Type [$F(3, 177) = 5.33$, $p = .002$, $\eta^2 = 0.08$], reflecting stronger effects of Set Size for *Mixed* displays than for *Fixed* displays.

For *Fixed* displays, there was no cost incurred by adding a face [FF, M = 734 ms, SE = 10 ms; FFF, M = 732 ms, SE = 11 ms; $F(1, 236) = 0.02$, $p = .889$, $\eta^2 = 0.00$] or a non-face [NN, M = 785 ms, SE = 11 ms; NNN, M = 775 ms, SE = 10 ms; $F(1, 236) = 0.76$, $p = .383$, $\eta^2 = 0.00$]. However, for *Mixed* displays, there was a significant cost to adding either a face [NF, M = 788 ms, SE = 10 ms; NFF, M = 831, SE = 10 ms; $F(1, 236) = 13.65$, $p < .001$, $\eta^2 = 0.05$] or a non-face [FN, M = 766 ms, SE = 7 ms; FNN, M = 803 ms, SE = 8 ms; $F(1, 236) = 10.55$, $p = .001$, $\eta^2 = 0.04$]. The effect of Display Type was significant for Set Size *Two* [$F(3, 354) = 5.49$, $p = .001$, $\eta^2 = 0.04$] and also for Set Size *Three* [$F(3, 354) = 15.90$, $p < .001$, $\eta^2 = 0.12$].

Data from the *Fixed* conditions were submitted to separate *t*-tests to compare detection of faces and non-faces at each Set Size. Detection was

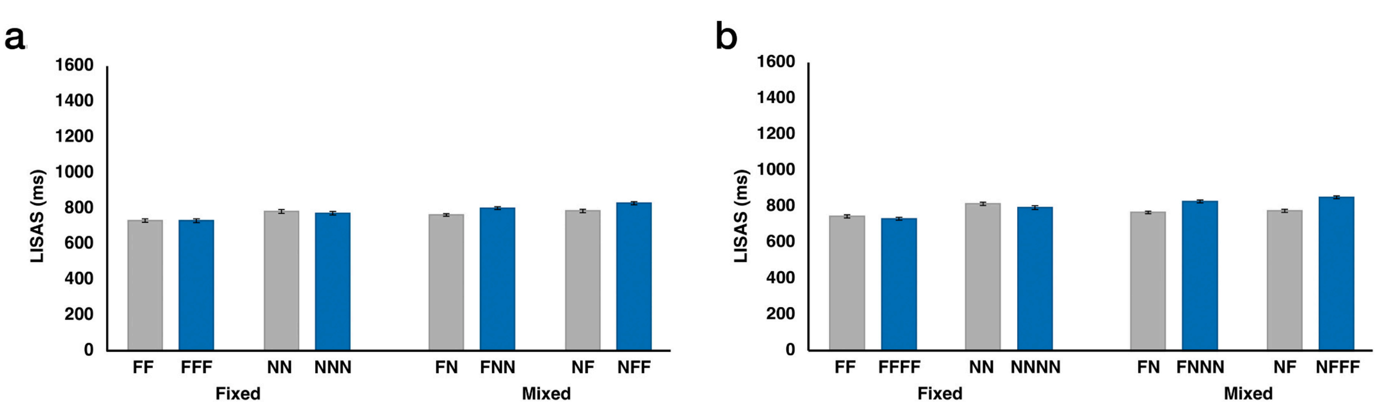

Fig. 2. Mean linear integrated speed-accuracy scores (LISAS) for each condition in (a) Experiment 1 and (b) Experiment 2. Lower scores indicate better efficiency. F denotes face, N denotes non-face (scrambled faces in Experiments 1 & 2). Error bars show within-subjects standard error (Cousineau, 2005).

significantly more efficient for faces than for non-faces at Set Size *Two* [t (59) = −2.97, p = .004, d = 0.38] and at Set Size *Three* [t (59) = −2.43, p = .018, d = 0.31].

Analysing responses to *Fixed* displays allowed us to compare detection efficiency for *Two* versus *Three* items of the same category (all faces or all non-faces). This comparison revealed no evidence of capacity limits, in the sense that there was no effect of set size: adding an extra item incurred no efficiency cost. However, at both set sizes, detection was more efficient for faces than for non-faces. The next experiment introduces a stronger manipulation of set size.

3. Experiment 2

To amplify possible effects of set size in this task, we next doubled the magnitude of the set size manipulation. Instead of adding one extra item to the displays (as in Experiment 1), we now added two extra items to the displays. We reasoned that doubling the strength of the set size manipulation should double the size of any latent performance costs.

3.1. Methods

3.1.1. Participants

Eighty new participants were recruited through Prolific and completed the experiment in exchange for a small payment. Twenty participants were excluded due to failed attention checks (2 or more within the same block, or 3 across the whole experiment) or slow responses (>2.5 SD from the group mean). The final sample (N = 60) comprised 21 females and 39 males (age range 18–49; M = 26.10, SD = 7.45).

3.1.2. Design and stimuli

The stimuli and catch trials were the same as in Experiment 1 but were now presented in displays of either two items or four items. A total of 68 faces and 68 non-faces were required for each of the Set Size *Two* conditions, and a total of 136 faces and 136 non-faces were required for each of the Set Size *Four* conditions. In each display, the selected items were randomly allocated to four predetermined locations that formed a square around central fixation (nearest contours ~95 pixels; see Fig. 1b).

3.1.3. Procedure

The procedure was the same as in Experiment 1, except that we increased the number of experimental trials. Participants now completed 4 experimental blocks, each consisting of 64 experimental trials (8 trials per conditions plus 3 catch trials) in a random order. The entire experiment took approximately 10 min to complete.

3.2. Results and discussion

Overall accuracy for the Fixed/Mixed judgements was 95%, confirming that participants could distinguish between face and non-face stimuli. Trials with reaction times below 150 ms or above 3000 ms were excluded from analysis (1.24% of all trials).

Fig. 2b summarises Linear Integrated Speed-Accuracy Scores for each condition in Experiment 2. LISAS data were submitted to a two-way ANOVA with the repeated measures factors of Set Size (*Two*, *Four*) and Display Type (*Fixed*, *Mixed*). This analysis revealed a significant main effect of Set Size, with more efficient detection for *Two* items (M = 762 ms, SE = 8 ms) than for *Four* items (M = 786 ms, SE = 9 ms) overall [F (1, 59) = 15.36, $p < .001$, $\eta^2 = 0.21$], and a significant main effect of display type, with more efficient detection for *Fixed* displays (M = 756 ms, SE = 8 ms) than for *Mixed* displays (M = 793 ms, SE = 8 ms) overall [F (3, 177) = 25.44, $p < .001$, $\eta^2 = 0.30$]. There was also a significant interaction between Set Size and Display Type [F (3, 177) = 18.97, $p < .001$, $\eta^2 = 0.26$], reflecting stronger effects of Set Size for *Mixed* displays than for *Fixed* displays.

For *Fixed* displays, there was no cost incurred by adding two extra faces [FF, M = 728 ms, SE = 8 ms; FFFF, M = 715 ms, SE =7 ms; F (1, 236) = 1.37, p = .243, $\eta^2 = 0.01$], or two extra non-faces [NN, M = 800 ms, SE = 8 ms; NNNN, M = 779 ms, SE = 10 ms; F (1, 236) = 3.34, p = .069, $\eta^2 = 0.01$]. However, for *Mixed* displays, there was a significant cost to adding either two faces [NF, M = 764 ms, SE = 9 ms; NFFF, M = 838, SE = 8 ms; F (1, 236) = 41.06, $p < .001$, $\eta^2 = 0.15$] or two non-faces [FN, M = 756 ms, SE = 6 ms; FNNN, M = 814 ms, SE = 7 ms; F (1, 236) = 25.86 $p < .001$, $\eta^2 = 0.15$]. The effect of Display Type was significant for Set Size *Two* [F (3, 354) = 10.74, $p < .001$, $\eta^2 = 0.08$] and also for Set Size *Four* [F (3, 354) = 35.14, $p < .001$, $\eta^2 = 0.23$].

Data from the *Fixed* conditions were submitted to separate t-tests to compare detection of faces and non-faces at each set size. Detection was significantly more efficient for faces than for non-faces at Set Size *Two*, [t (59) = −5.12, $p < .001$, d = 0.66] and at Set Size *Four* [t (59) = −4.71, $p < .001$, d = 0.61].

As in Experiment 1, comparing detection efficiency for items of the same category revealed no evidence of capacity limits, despite the fact that the magnitude of the set size manipulation was now doubled. Apparently, acquiring four items was no less efficient than acquiring two. We again found that detection was more efficient for faces than for non-faces. Given that stimulus–template match depends on properties of the stimulus, we next asked what differences between faces and non-faces are required for efficient detection of multiple faces.

4. Experiment 3

The high detection efficiency for faces in Experiments 1 and 2 cannot have been due to their low-level visual energies, given the different result for the phase-shifted faces. However, phase-shifted faces do not control for contour and structural information in the intact face. In the next experiment, we used inverted faces as the comparison stimuli instead. For consistency across experiments, we refer to these inverted face stimuli as 'non-faces'. Given that inverted faces are identical to upright faces in every respect except orientation, we expected the visual distinction between face and non-face stimuli to be less clear, potentially reducing task efficiency.

4.1. Methods

4.1.1. Participants

Seventy-three participants, who were recruited online via Prolific, completed the experiment in exchange for a small payment. Thirteen participants were excluded due to failed attention checks (2 or more within the same block, or 3 across the whole experiment) or slow responses (>2.5 SD from the group mean). The final sample (N = 60) comprised 26 females and 34 males (age range 19–61; M = 30.22; SD = 9.11).

4.1.2. Design and stimuli

The design was the same as for Experiment 1 except that the intact and scrambled faces were replaced with upright and inverted faces. The upright faces were taken from the original bank of 288 faces and segmented from the background using the InterFace software package (Kramer, Jenkins, & Burton, 2017). The resulting images were resized to 570 pixels high × 380 pixels wide. The non-face stimuli were created by rotating the upright faces 180° in the picture plane. As in Experiment 1, stimuli were combined to create *Fixed* and *Mixed* displays of Set Size *Two* and Set Size *Three*. Example displays are shown in Fig. 3.

Catch trials for each condition were again created for use as attention checks. These catch trials were constructed in a similar manner and used the same spatial layout as the experimental trials. White circles containing upward- or downward-pointing black arrows were used in place of upright or inverted faces, respectively.

4.1.3. Procedure

The procedure was the same as in Experiment 1 and took

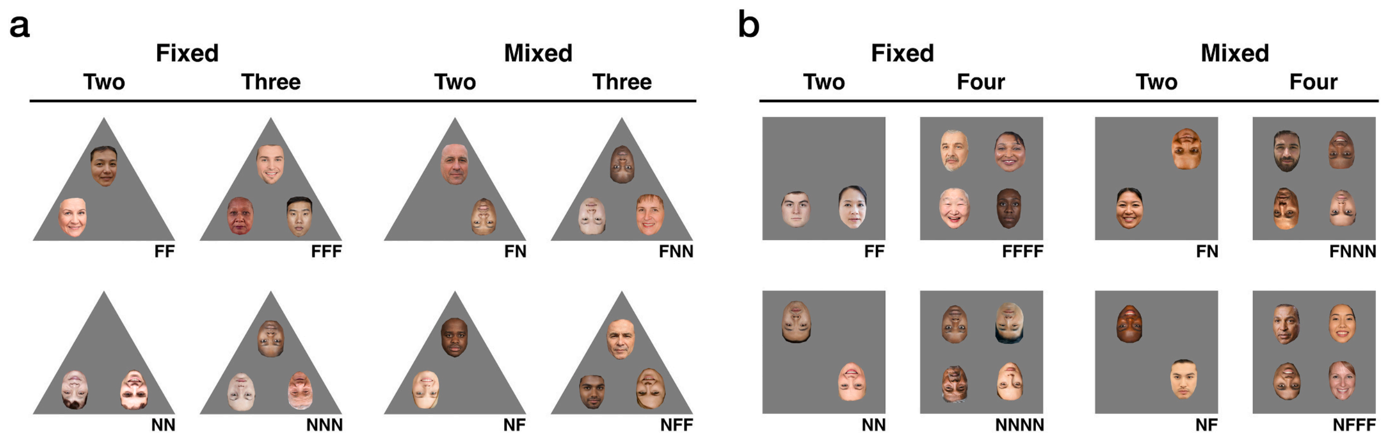

Fig. 3. Example displays for each condition of (a) Experiment 3 and (b) Experiment 4. Fixed displays contained one type of stimulus (all faces or all non-faces). Mixed displays contained both types. Numbers refer to set sizes. F denotes face, N denotes non-face (inverted faces in Experiments 3 & 4). Triangle and square segments are for visualisation only. In the actual experiments, the grey background filled the whole screen.

approximately 10 min to complete.

4.2. Results and discussion

Overall accuracy for the Fixed/Mixed judgements was 94%, again confirming that participants could distinguish between face and non-face stimuli, despite their increased similarity in this experiment. Trials with reaction times below 150 ms or above 3000 ms were excluded from analysis (1.6% of all trials).

Fig. 4a summarises Linear Integrated Speed-Accuracy Scores for each condition in Experiment 3. LISAS data were submitted to a two-way ANOVA with the repeated measures factors of Set Size (*Two*, *Three*) and Display Type (*Fixed*, *Mixed*) This analysis revealed a significant main effect of Set Size, with more efficient detection for *Two* items (M = 991 ms, SE = 12 ms) than for *Three* items (M = 1092 ms, SE = 14 ms) overall [F (1, 59) = 127.29, $p < .001$, $\eta^2 = 0.68$], and a significant main effect of display type, with more efficient detection for *Fixed* displays (M = 985 ms, SE = 15 ms) than for *Mixed* displays (M = 1098 ms, SE = 13 ms) overall [F (3, 177) = 31.68, $p < .001$, $\eta^2 = 0.39$]. There was also a significant interaction between Set Size and Display Type [F (3, 177) = 27.65, $p < .001$, $\eta^2 = 0.32$], reflecting stronger effects of Set Size for *Mixed* displays than for *Fixed* displays.

For *Fixed* displays, there was a significant cost incurred by adding an extra face [FF M = 915 ms, SE = 12 ms; FFF, M = 952 ms, SE =14 ms; F (1, 236) = 4.89, $p = .028$, $\eta^2 = 0.03$] or an extra non-face [NN, M = 1018 ms, SE = 18 ms; NNN, M = 1056 ms, SE = 17 ms; F (1, 236) = 4.99, $p = .028$, $\eta^2 = 0.02$].

For *Mixed* displays, there was also a significant cost to adding an extra face [NF, M = 1022 ms, SE = 11 ms; NFF, M = 1131 ms, SE = 12 ms; F (1, 236) = 41,60, $p < .001$, $\eta^2 = 0.15$] or an extra non-face [FN, M = 1009 ms, SE = 11 ms; FNN, M = 1230 ms, SE = 16 ms; – F (1, 236) = 171.91, $p < .001$, $\eta^2 = 0.42$]. The effect of Display Type was significant for Set Size *Two* [F (3, 354) = 11.04, $p < .001$, $\eta^2 = 0.09$] and also for Set Size *Three* [F (3, 354) = 58.77, $p < .001$, $\eta^2 = 0.33$].

Data from the *Fixed* conditions were submitted to separate *t*-tests to compare detection of faces and non-faces at each set size. Detection was significantly more efficient for faces than for non-faces at Set Size *Two*, [*t* (59) = −4.33, $p < .001$, $d = 0.56$], and at Set Size *Three* [*t* (59) = −4.25, $p < .001$, $d = 0.55$].

Analysis of the *Fixed* conditions again allowed us to compare detection efficiency for *Two* versus *Three* items of the same category (all faces or all non-faces). Unlike Experiment 1, this comparison revealed clear evidence of capacity limits, in the sense that there was a significant effect of set size: adding an extra item incurred a substantial efficiency cost, as expected from the reduced distinction between face and non-face stimuli. In keeping with the preceding experiments, detection was again more efficient for faces than for non-faces. This face advantage was seen not only in the *Fixed* conditions, but also in the *Mixed* conditions, where adding a face to a display incurred a smaller cost than adding a non-face.

5. Experiment 4

To better understand the effects of set size seen in Experiment 3, we doubled the magnitude of the set size manipulation (similar to Experiment 2). If performance is capacity-limited for these new stimuli, such

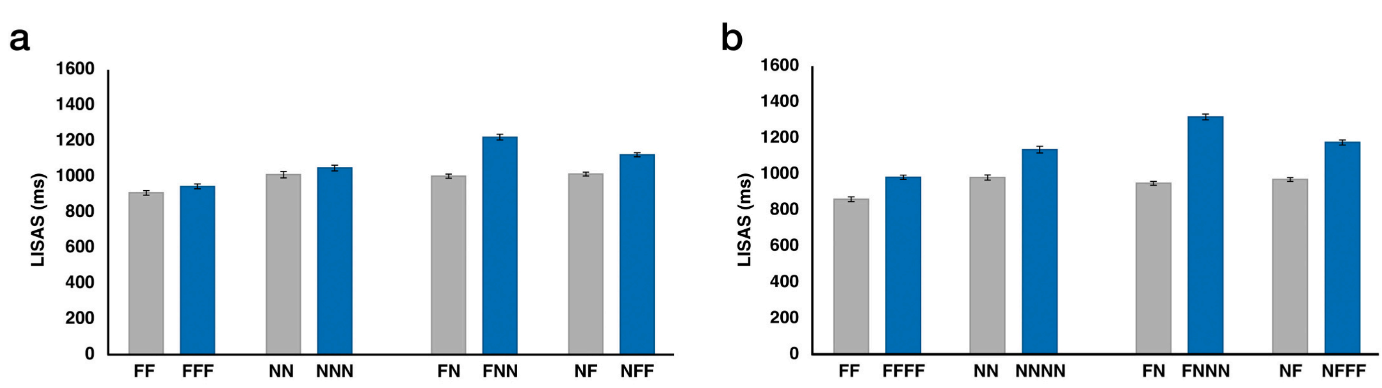

Fig. 4. Mean linear integrated speed-accuracy scores (LISAS) for each condition in (a) Experiment 3 and (b) Experiment 4. Lower scores indicate better efficiency. F denotes face, N denotes non-face (inverted faces in Experiments 3 & 4). Error bars show within-subjects standard error (Cousineau, 2005).

that each extra item incurs its own performance cost, then doubling the number of extra items (from 1 to 2) should double the cost.

5.1. Methods

5.1.1. Participants

Sixty-eight new participants were recruited through Prolific and completed the experiment in exchange for a small payment. Eight participants were excluded due to failed attention checks (2 or more within the same block, or 3 across the whole experiment) or slow responses (>2.5 SD from the group mean). The final sample ($N = 60$) comprised 23 females and 37 males (age range 18–56; M = 27.25; SD = 8.93).

5.1.2. Design and stimuli

The design, stimuli, and catch trials, were the same as in Experiment 3, except that the display items comprised either two or four items that were randomly allocated to four predetermined locations that formed a square around central fixation (see Fig. 3b).

5.1.3. Procedure

The procedure was the same as in Experiment 2 and took approximately 10 min to complete.

5.2. Results and discussion

Overall accuracy for the Fixed/Mixed judgements was 92%, again confirming that participants could distinguish between face and non-face stimuli. Trials with reaction times below 150 ms or above 3000 ms were excluded from analysis (0.97% of all trials).

Fig. 4b summarises Linear Integrated Speed-Accuracy Scores for each condition in Experiment 4. LISAS data were submitted to a two-way ANOVA with the repeated measures factors of Set Size (*Two*, *Four*) and Display Type (*Fixed*, *Mixed*). This analysis revealed a significant main effect of Set Size, with more efficient detection for *Two* items (M = 981 ms, SE = 14 ms) than for *Four* items (M = 1202 ms, SE = 15 ms) overall [$F (1, 59) = 313.66, p < .001, \eta^2 = 0.84$], and a significant main effect of display type, with more efficient detection for *Fixed* displays (M = 1031 ms, SE = 15 ms) than for *Mixed* displays (M = 1132 ms, SE = 13 ms) overall [$F (3, 177) = 57.87, p < .001, \eta^2 = 0.50$]. There was also a significant interaction between Set Size and Display Type [$F (3, 177) = 46.62, p < .001, \eta^2 = 0.44$], reflecting stronger effects of Set Size for *Mixed* displays than for *Fixed* displays.

For *Fixed* displays, there was a significant cost incurred by adding two extra faces [FF, M = 895 ms, SE = 14 ms; FFFF, M = 1021 ms, SE =13 ms; $F (1, 236) = 44.00, p < .001, \eta^2 = 0.16$], or two extra non-faces [NN, M = 1023 ms, SE = 14 ms; NNNN, M = 1184 ms, SE = 119 ms; $F (1, 236) = 70.82, p < .001, \eta^2 = 0.23$].

Similarly, for *Mixed* displays, there was a significant cost to adding two extra faces [NF, M = 1014 ms, SE = 11 ms; NFFF, M = 1227 ms, SE = 14 ms; $F (1, 236) = 124.93, p < .001, \eta^2 = 0.35$] or two extra non-faces [FN, M = 991 ms, SE = 11 ms; FNNN, M = 1375 ms, SE = 17 ms; $F (1, 236) = 403.01, p < .001, \eta^2 = 0.63$]. The effect of Display Type was significant for Set Size *Two* [$F (3, 354) = 15.29, p < .001, \eta^2 = 0.11$] and for Set Size *Four* [$F (3, 354) = 93.51, p < .001, \eta^2 = 0.44$].

Data from the *Fixed* conditions were submitted to separate *t*-tests to compare detection of faces and non-faces at each set size. Detection was significantly more efficient for faces than for non-faces at Set Size *Two*, [$t (59) = -6.34, p < .001, d = 0.82$] and at Set Size *Four* [$t (59) = -6.41, p < .001, d = 0.83$].

As with Experiment 3, the results of Experiment 4 are consistent with capacity-limited performance in which each additional display item contributes to performance costs. Detection was again more efficient for faces than for non-faces. This face advantage was observed across all experimental conditions.

6. General discussion

Our experiments reveal at least four principles of multiple face detection. First and foremost, we show that viewers can capture additional faces at no extra cost. Second, cost-free capture is contingent on visual context. Third, this facility is not specific to faces. Fourth, it is efficient for faces. We address each of these points in turn.

Can viewers know that more than one face is present? Yes. Fig. 5 summarises the cost per additional item for fixed/mixed judgements across experiments. Experiment 1 shows that adding an item to the display incurred no efficiency cost. Even when the number of items was doubled from two to four (Experiment 2), we found no impact on efficiency. In none, of these cases was the per-item cost even numerically positive. We conclude that the visual system can acquire multiple faces concurrently, at least over the range of 2–4 faces tested here. This finding is consistent with the ERP observation that multiple faces can enhance the N170 (Puce et al., 2013). However, the current task allows us to draw more specific conclusions. Unlike a simple numerosity task, the fixed/mixed task used here required participants to discern whether or not the seen items were intact upright faces.

The observed multi-item capacity at the detection stage of face processing contrasts with surprisingly strict capacity limits seen for later stages of face processing. Response competition experiments requiring judgements of sex, eye direction, and semantic information have repeatedly found that processing one face precludes processing another face (Bindemann et al., 2007; Bindemann, Burton, et al., 2005; Burton & Bindemann, 2009; Jenkins et al., 2003). The cognitively earlier task of face detection apparently evades this strict limit of one. Taken together, these findings help to locate the putative bottleneck in face processing. We suggest that a processing bottleneck occurs *after* the detection step, such that coarse face/non-face discriminations may be conducted in parallel, but before further face information is extracted, such that finer discriminations among faces must be conducted in series. Future experiments could modify the method described here to test the upper bound of multiple face detection—in particular, whether detection capacity exceeds our maximum of four presented items. We expect that there will be some limit to the number of faces that can be detected concurrently, not least because overall visual bandwidth is limited. Establishing these upper limits for face detection will require careful experimentation, as increasing the number of display items necessitates increasing eccentricity, increasing crowding, or reducing item size, all of which can affect general visual discrimination. One interesting possibility relates to *subitizing*—rapid and accurate enumeration of up to 4 items (Kaufman, Lord, Reese, & Volkmann, 1949; Piazza, Fumarola, Chinello, & Melcher, 2011; Trick & Pylyshyn, 1994). Although classic demonstrations of subitizing relied on simple visual objects (e.g. dots) as stimuli, more recent work has established subitizing-like phenomena for complex objects, including human figures (Railo et al., 2016). The fixed/mixed task is clearly different from subitizing, as it requires participants to know whether or not display items are of the same type, rather than just enumerating them. Even so, the 4-item span associated with subitizing provides theoretical motivation to test larger set sizes in future experiments. For now, equivalent detection efficiency over the range of 2 to 4 items shows that face detection is not subject to a strict capacity limit of just one face.

The comparison between Experiments 1 & 2 (in which the non-faces were scrambled faces) and Experiments 3 & 4 (in which the non-faces were inverted faces) underscores the importance of visual context in determining task performance (see Fig. 5). Although multiple face detection *can* proceed in parallel (Experiments 1 & 2), whether or not it actually *does* proceed in parallel depends on other factors—in this case, the nature of other items in the display (Experiments 3 & 4). This contingency is useful, as it suggests a way to probe what counts as a face to the visual system. Establishing that scrambled faces are not faces incurred no cognitive cost in this task, implying that an appropriate distribution of visual energies is not itself sufficient to match the face

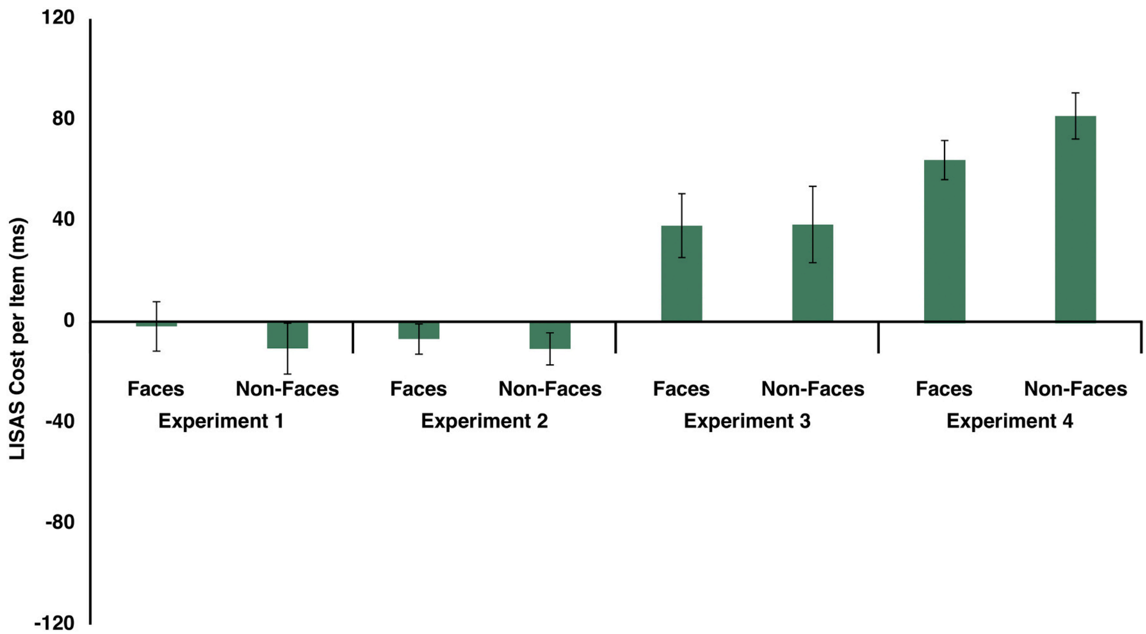

Fig. 5. Summary of cost per additional item for Fixed conditions (face and non-face) in Experiments 1–4.

template. In contrast, establishing that inverted faces are not faces did incur a cognitive cost, implying that the spatial organisation of those visual energies makes a meaningful difference. We suggest that an inverted face matches (or partially matches) the face detection template whereas a scrambled face does not. As a result, upright and inverted faces take additional cognitive resources to sort out, apparently one at a time.

This interpretation fits with the broader notion that discriminations among stimuli that activate face processing (e.g. identification; extraction of social signals) are serial in nature. It also suggests a candidate behavioural marker of face template matching. If stimuli that give rise to serial processing in this task engage the template, and stimuli that give rise to parallel processing do not, it should be possible to characterise the 'receptive field' of the face template by varying the non-face stimuli in this task. Further iterations of the present task, with different kinds of carefully-designed distractor items, could reveal what counts (and does not count) as a face to the visual system.

It may seem counterintuitive to emphasise non-face visual content as a determinant of face perception, especially as so much previous work emphasises facial appearance. However, the distinction between face and non-face is key to the early perceptual step of face detection. From this perspective, we should expect performance to be determined as much by the rest of the visual environment as it is by faces themselves. The more closely visual properties of the environment resemble visual properties of faces (cf. scrambled faces in Experiments 1 & 2, inverted faces in Experiments 3 & 4), the more demanding the face/non-face discrimination becomes (Duncan & Humphreys, 1989; Lewis & Edmonds, 2005). This basic insight suggests that it will be difficult to generalise from detection experiments based on isolated faces only to face detection in the real world. It also suggests a principled means to reconcile seemingly discrepant findings in the literature. Visual search studies that have reported 'pop-out' for a target face have avoided presenting other face-like information in their displays (e.g. Lewis & Edmonds, 2005, Experiment 1). Those that found no pop-out did present other face-like information (notably inverted faces, e.g. Lewis & Edmonds, 2005, Experiment 2; Nothdurft, 1993; Brown et al., 1997). This distinction among studies of visual search for faces echoes more general findings in visual search. Search is most efficient when targets and distractors are dissimilar and displays contain homogeneous distractors; search becomes less efficient when target–distractor similarity increases irrespective of display heterogeneity (Roper, Cosman, & Vecera, 2013). However, all of those studies relied specifically on a visual search task, in which the experimenters define faces as targets, viewers must localise display items, and a maximum of one target face is present. The current experiments emphasise the relation between face and non-face material in a very different task, in which no target category is defined, localisation is not required, and multiple faces are acquired from the visual environment simultaneously.

Our final two points concern whether or not faces are 'special' in this situation. We note that response patterns across experiments were qualitatively similar for face and non-face displays. In Experiments 1 and 2, comparison of fixed conditions revealed no evidence of capacity limits for either stimulus category, in that adding extra items incurred no efficiency cost. In Experiments 3 and 4, the same comparison revealed set-size costs for both stimulus categories. As such, we make no claims concerning *qualitative* differences between multiple face detection and multiple stimulus detection generally. However, *quantitative* differences between categories were both clear and consistent. Across all four experiments, responses to faces were more efficient than responses to non-faces. This apparent face advantage accords with previous studies of face detection. For example, detection of upright intact faces has been shown to be more efficient than detection of other objects, pareidolic faces, and even faces with rearranged internal features (Crouzet et al., 2010; Keys, Taubert, & Wardle, 2021; Purcell & Stewart, 1986, 1988; Stein et al., 2012). In our view, there are many possible explanations for this apparent face advantage. The current experiments were not designed to disentangle them. Instead, we conclude that multiple faces can be detected concurrently, implying that the bottleneck in face processing follows the detection step, rather than preceding or coinciding with detection. Whether multiple faces actually are detected concurrently in a particular situation can depend on other aspects of the visual scene.

CRediT authorship contribution statement

Rana Qarooni: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft. **Jonathan Prunty:** Conceptualization, Writing – review & editing. **Markus Bindemann:** Conceptualization, Writing – review & editing. **Rob Jenkins:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Data availability

Data for these experiments are publicly available on the Open Science Framework at https://osf.io/v4c9e/.

Acknowledgements

This work was funded by a research grant from the Leverhulme Trust (RPG-2019-085) to Markus Bindemann and Rob Jenkins.

Appendix A. Supplementary data

Statistical analysis of accuracy and reaction time measures for Experiments 1–4.

References

Amso, D., Haas, S., & Markant, J. (2014). An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PLoS One, 9*(1), Article e85701. https://doi.org/10.1371/JOURNAL.PONE.0085701

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: an online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bindemann, M., & Burton, A. M. (2009). The role of color in human face detection. *Cognitive Science, 33*(6), 1144–1156. https://doi.org/10.1111/j.1551-6709.2009.01035.x

Bindemann, M., Burton, A. M., Hooge, I. T. C., Jenkins, R., & De Haan, E. H. F. (2005). Faces retain attention. In *, Vol. 12, 6. Psychonomic bulletin and review* (pp. 1048–1053). Psychonomic Society Inc. https://doi.org/10.3758/BF03206442.

Bindemann, M., Burton, A., & Jenkins, R. (2005). Capacity limits for face processing. *Cognition, 98*(2), 177–197. https://doi.org/10.1016/j.cognition.2004.11.004

Bindemann, M., Jenkins, R., & Burton, A. M. (2007). A bottleneck in face identification: Repetition priming from flanker images. *Experimental Psychology, 54*(3), 192–201. https://doi.org/10.1027/1618-3169.54.3.192

Bindemann, M., & Lewis, M. B. (2013). Face detection differs from categorization: Evidence from visual search in natural scenes. *Psychonomic Bulletin and Review, 20* (6), 1140–1145. https://doi.org/10.3758/s13423-013-0445-9

Brown, V., Huey, D., & Findlay, J. M. (1997). Face detection in peripheral vision: Do faces pop out? *Perception, 26*(12), 1555–1570. https://doi.org/10.1068/P261555

Bruckmaier, M., Tachtsidis, I., Phan, P., & Lavie, N. (2020). Attention and capacity limits in perception: A cellular metabolism account. *Journal of Neuroscience, 40*(35), 6801–6811. https://doi.org/10.1523/JNEUROSCI.2368-19.2020

Burton, A. M., & Bindemann, M. (2009). The role of view in human face detection. *Vision Research, 49*(15), 2026–2036. https://doi.org/10.1016/j.visres.2009.05.012

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorial in Quantitative Methods for Psychology, 1*(1), 42–45. https://doi.org/10.20982/TQMP.01.1.P042

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision, 10*(4), 1–17. https://doi.org/10.1167/10.4.16

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity the efficiency of visual selection. *Psychological Review, 96*(3).

Garner, K. G., Bowman, H., & Raymond, J. E. (2021). Incentive value and spatial certainty combine additively to determine visual priorities. *Attention, Perception, & Psychophysics, 83*(1), 173–186. https://doi.org/10.3758/S13414-020-02124-W

Hjelmås, E., & Low, B. K. (2001). Face detection: A survey. *Computer Vision and Image Understanding, 83*(3), 236–274. https://doi.org/10.1006/cviu.2001.0921

Honey, C., Kirchner, H., & VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *Journal of Vision, 8(12)*(9), 1–12. https://doi.org/10.1167/8.12.9

Jenkins, R., Lavie, N., & Driver, J. (2003). Ignoring famous faces: Category-specific dilution of distractor interference. *Perception & Psychophysics, 65*(2), 298–309. https://doi.org/10.3758/BF03194801

Karras, Nvidia, et al. (2018). Generated Photos | Unique, worry-free model photos. https://generated.photos/.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology, 62*(4), 498–525.

Keys, R. T., Taubert, J., & Wardle, S. G. (2021). A visual search advantage for illusory faces in objects. *Attention, Perception, & Psychophysics, 83*(5), 1942–1953.

Kramer, R. S. S., Jenkins, R., & Burton, A. M. (2017). InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods, 49*(6), 2002–2011. https://doi.org/10.3758/s13428-016-0837-7

Kuehn, S. M., & Jolicoeur, P. (1994). Impact of quality of the image, orientation, and similarity of the stimuli on visual search for faces. *Perception, 23*(1), 95–122. https://doi.org/10.1068/p230095

Kumar, A., Kaur, A., & Kumar, M. (2018). Face detection techniques: a review. *Artificial Intelligence Review, 52*(2), 927–948. https://doi.org/10.1007/S10462-018-9650-2

Lavie, N., & De Fockert, J. W. (2003). Contrasting effects of sensory limits and capacity limits in visual selective attention. *Perception & Psychophysics, 65*(2), 202–212. https://doi.org/10.3758/BF03194795

Lewis, M. B., & Edmonds, A. J. (2005). Searching for faces in scrambled scenes. *Visual Cognition, 12*(7), 1309–1336. https://doi.org/10.1080/13506280444000535

Lewis, M. B., & Ellis, H. D. (2003). How we detect a face: A survey of psychological evidence. *International Journal of Imaging Systems and Technology, 13*(1), 3–7. https://doi.org/10.1002/ima.10040

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*(1), 44–64. https://doi.org/10.1016/0010-0285(75)90004-3

Nothdurft, H.-C. (1993). Faces and facial expressions do not pop out. In. *Perception, 22*.

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

Pongakkasira, K., & Bindemann, M. (2015). The shape of the face template: Geometric distortions of faces and their detection in natural scenes. *Vision Research, 109*(Part A), 99–106. https://doi.org/10.1016/j.visres.2015.02.008

Puce, A., McNeely, M. E., Berrebi, M. E., Thompson, J. C., Hardee, J., & Brefczynski-Lewis, J. (2013). Multiple faces elicit augmented neural activity. *Frontiers in Human Neuroscience, 7*(MAY), 282. https://doi.org/10.3389/fnhum.2013.00282

Purcell, D. G., & Stewart, A. L. (1986). The face-detection effect. *Bulletin of the Psychonomic Society, 24*(2), 118–120.

Purcell, D. G., & Stewart, A. L. (1988). The face-detection effect: Configuration enhances detection. *Perception & Psychophysics, 43*(4). https://doi.org/10.3758/BF03208806

Robertson, D. J., Jenkins, R., & Burton, A. M. (2017). Face detection dissociates from face identification. *Visual Cognition, 25*(7–8), 740–748. https://doi.org/10.1080/13506285.2017.1327465

Roper, Z. J., Cosman, J. D., & Vecera, S. P. (2013). Perceptual load corresponds with factors known to influence visual search. *Journal of Experimental Psychology: Human Perception and Performance, 39*(5), 1340–1351.

Simpson, E. A., Maylott, S. E., Leonard, K., Lazo, R. J., & Jakobsen, K. V. (2019). Face detection in infants and adults: Effects of orientation and color. *Journal of Experimental Child Psychology, 186*, 17–32. https://doi.org/10.1016/j.jecp.2019.05.001

Stein, T., Sterzer, P., & Peelen, M. V. (2012). Privileged detection of conspecifics: Evidence from inversion effects during continuous flash suppression. *Cognition, 125* (1), 64–79. https://doi.org/10.1016/j.cognition.2012.06.005

Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., & Heagins, T. B. (2016). The MR2: A multi-racial, mega-resolution database of facial stimuli. *Behavior Research Methods, 48*(3), 1197–1204. https://doi.org/10.3758/s13428-015-0641-9

Thoma, V., & Lavie, N. (2013). Perceptual load effects on processing distractor faces indicate face-specific capacity limits. *Visual Cognition, 21*(8), 1053–1076. https://doi.org/10.1080/13506285.2013.853717

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

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.

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience, 31*(1), 411–437. https://doi.org/10.1146/annurev.neuro.30.051606.094238

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods, 49*(2), 653–673. https://doi.org/10.3758/s13428-016-0721-5

Vandierendonck, A. (2018). Further tests of the utility of integrated speed-accuracy measures in task switching. *Journal of Cognition, 1*(1). https://doi.org/10.5334/joc.6

Vandierendonck, A. (2021). On the utility of integrated speed-accuracy measures when speed-accuracy trade-off is present. *Journal of Cognition, 4*(1), 1–26. https://doi.org/10.5334/joc.154

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*(6), 495–501. https://doi.org/10.1038/nrn1411