



# How do we react to cluttered displays? Evidence from the first seconds of visual search in websites

Malk Kanaan & Nadine Marie Moacdieh

**To cite this article:** Malk Kanaan & Nadine Marie Moacdieh (2021) How do we react to cluttered displays? Evidence from the first seconds of visual search in websites, *Ergonomics*, 64:11, 1452-1464, DOI: [10.1080/00140139.2021.1927200](https://doi.org/10.1080/00140139.2021.1927200)

**To link to this article:** <https://doi.org/10.1080/00140139.2021.1927200>



Published online: 01 Jun 2021.



Submit your article to this journal [↗](#)



Article views: 636



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

ARTICLE



## How do we react to cluttered displays? Evidence from the first seconds of visual search in websites

Malk Kanaan and Nadine Marie Moacdieh 

Department of Industrial Engineering and Management, American University of Beirut, Beirut, Lebanon

### ABSTRACT

Display clutter is known to degrade search performance and lead to differences in eye movement measures in different contexts. The goal of this study was to determine whether these differences in eye movements could be detected in the first few seconds of a search task using a realistic display, both with or without time pressure. Participants were asked to search for image or word targets in 40 website screenshots. Time pressure was introduced for half the trials. Clutter algorithms were used to classify the websites as low- or high-clutter. Performance, subjective, and eye-tracking metrics were collected. Results showed that people's attention allocation within the first 3 s of search is different when viewing high-clutter websites. In particular, people's spread of attention was larger in high-clutter websites. The results can be used to detect whether a person is struggling with clutter early on after they view a display.

**Practitioner summary:** Eye-tracking metrics showed that people react differently to a cluttered website in a variety of conditions. These differences were evident within the first 3 s of the search. The eye-tracking metrics identified can be used to detect people struggling with clutter as soon as they look at a website.

**Abbreviations:** RT: response time; FC: feature congestion; SE: subband entropy; ED: edge density; NNI: nearest neighbor index; TLX: task load index

### ARTICLE HISTORY

Received 27 October 2019  
Accepted 3 May 2021

### KEYWORDS

Display clutter; visual search; eye tracking; websites; interface design

### Introduction and motivation

Display clutter, which will be referred to as clutter for simplicity, can be defined as the presence of high data density, poor display organisation, and an abundance of irrelevant information that leads to performance decrements (Moacdieh and Sarter 2015a). Clutter has been shown to negatively affect performance, particularly in the context of visual search (e.g. Neider and Zelinsky 2011). More specifically, display clutter can degrade monitoring and signal/change detection (Schons and Wickens 1993), delay visual search (Henderson, Chanceaux, and Smith 2009; Neider and Zelinsky 2011) and negatively affect object recognition (Bravo and Farid 2008). It follows that a display designer would want to know how much clutter is in a certain display in order to ensure that users can quickly and effectively find their search target.

The difficulty in establishing a reliable measure of clutter is the presence of numerous factors that can affect visual search (Wolfe and Horowitz 2017). In general, these factors can be divided into those that are

bottom-up (i.e. display-based), such as target-background colour similarity, and top-down (i.e. user-based), such as the user's goals and experience (Wolfe et al. 2011). The interaction between these factors is what determines response time (RT). In other words, what may be 'cluttered' to one user (i.e. longer RT) may not be to another user.

The main challenges to clutter measurement are then twofold. The first challenge is to find a measure of clutter that would incorporate not just the number and organisation of items on a display, but also the user's reaction to these items. The second is to be able to capture this reaction in real-time and early on in the search process so that if the user is struggling, any possible display adjustments can be made to the display, and if not, the display is left unchanged. This ability to adjust *if needed* is a hallmark of adaptive, intelligent displays that can help rather than further overwhelm the user with information (Kaber et al. 2001). Three techniques that are commonly used for measuring clutter are image processing algorithms

(e.g. Bravo and Farid 2008; Pankok and Kaber 2018; Rosenholtz, Li, and Nakano 2007), performance assessment (e.g. Wickens et al. 2005), and subjective assessment (e.g. Kaufmann and Kaber 2010), but none of them can meet the desired challenges. A novel approach by Pankok and Kaber (2018) uses a combination of top-down and bottom-up factors, but that still cannot fully capture each person's reaction to clutter.

Eye-tracking is one approach that can address the limitations of the aforementioned techniques (Moacdieh and Sarter 2015b). However, not enough is known about eye tracking and clutter early on in the search process. The overall goal of this study is then to determine whether and how eye tracking can be used to detect differences in attention allocation to cluttered versus uncluttered real-life displays in the first few seconds of visual search.

### *Eye-tracking for clutter research*

Eye-tracking has been used to a good extent when it comes to understanding the effects of clutter (Beck, Lohrenz, and Trafton 2010; Fabrikant, Hespanha, and Hegarty 2010; Hegarty, De Leeuw, and Bonura 2008; Henderson, Chanceaux, and Smith 2009; Neider and Zelinsky 2011). For example, some studies showed a significant increase in the number of fixations in more cluttered aeronautical charts and websites (e.g. Beck, Lohrenz, and Trafton 2010; Beck et al. 2012).

However, despite this body of knowledge, far fewer studies have looked into how clutter affects eye movements in the first few seconds of search on a display. In general, it is known that, at a glance, people can categorise natural scenes (e.g. Ehinger and Rosenholtz 2016), classify the type of website, and identify menu bars (Jahanian, Keshvari, and Rosenholtz 2018). More specifically, in the context of clutter, Moacdieh and Sarter (2015b) investigated its effects on eye-tracking metrics in the context of electronic medical records (EMRs), and did not find any significant effects of clutter during the first 4 s of search. Moacdieh and Sarter (2017a) later found that three eye-tracking metrics calculated over the first 3 s of search can predict the presence of clutter (Moacdieh and Sarter 2017b). This study was done in the context of a simple graphics display. However, EMRs are a specialised type of display that only trained physicians can use, and the graphics display used was developed for the study and was highly controlled. Other types of displays that have been used in clutter research include maps (Beck, Lohrenz, and Trafton 2010; Rosenholtz, Li, and Nakano 2007), real-life images (Asher et al. 2013), or purposefully-developed

computer-generated images of buildings (e.g. Neider and Zelinsky 2011). However, these types of images are very similar in that they are largely image-based displays (i.e. with little to no words) that represent only a small subset of the types of displays people use.

Moreover, the targets that people have been asked to search for have often been artificial elements that are added to the display by the experimenters, such as Gabor patches (Rosenholtz, Li, and Nakano 2007), letters (Henderson, Chanceaux, and Smith 2009), or asterisks (Tuch et al. 2009). There is a need to study clutter in real-life displays – such as websites – that are commonly used and that contain both words and images as part of the display. Search for words can also be very different from the search for images (Paivio and Begg 1974), with the colours associated with images often making them easier to search for in a bottom-up fashion (Wolfe et al. 2011). Note that the idea is not to compare RT for images versus words, as there are too many factors to consider to make a direct comparison. Instead, the question here is whether eye tracking can be reliably used whether one is searching for an image or a word on a website.

### *The problem of time pressure*

Moreover, the effects of clutter may be exacerbated by factors such as stress, fatigue, and time pressure (Moacdieh and Sarter 2015b; Naylor 2010). It is certainly to a website designer's benefit to consider the fact that people might be in a rush when opening a website, and if they cannot find what they are looking for they might easily leave. Previous research has been inconsistent in terms of the interaction effects of clutter and time pressure (Moacdieh and Sarter 2015b, 2017a), with time pressure causing both shorter and longer RTs in high clutter, although the error rate generally increases. It has been shown in market research that time pressure can speed up the visual search (Pieters and Warlop 1999), although there is not enough in the literature to make a conclusion for websites. It has been shown that people can process elements of a website – including words – within 120 ms after being presented with a website screenshot (Jahanian, Keshvari, and Rosenholtz 2018). This suggests that people's top-down knowledge of a typical website's design helps identify important elements early on. It could be that this ability becomes more prominent under time pressure, as users try to speed up the process by relying on what they already know.

**Table 1.** A list of the eye tracking metrics used in this study.

Name	Explanation
<b>Spread metrics</b>	
Convex Hull area (pixels <sup>2</sup> )	The minimum convex area which contains the fixation points (Goldberg and Kotval 1999). This is calculated using the Matlab function <code>convHull</code> , with the X and Y positions of the fixation points as input. The maximum area of the screen is $1920 \times 1200 = 2.304 \times 10^6$ pixels <sup>2</sup> . A higher convex hull area indicates more spread of attention.
Spatial density	The number of grid cells containing gaze points divided by the total number of cells (Goldberg and Kotval 1999). A $20 \times 20$ grid of equally-sized cells ( $96 \times 60$ pixels per cell) was created to cover the whole screen. Similar to convex hull area, a higher spatial density would indicate a larger dispersion of attention.
Nearest neighbour index (NNI)	The ratio between (1) the average of the observed minimum distances between fixations and (2) the mean random distance between fixations expected if the distribution were random, calculated as half the square root of display area divided by the number of fixations (Di Nocera, Camilli, and Terenzi 2007). A higher NNI indicates more clutter
<b>Directness metrics</b>	
Scanpath length per second (pixels/s)	The sum of all the saccade lengths divided by the total time. Similar to mean saccade amplitude, a larger scanpath length indicates less directness.
Backtrack rate (/s)	A backtrack is defined as an angle between two saccades that is greater than $90^\circ$ , representing a change in the direction of the search (Goldberg and Kotval 1999). A larger backtrack rate indicates less directness.
Rate of transitions (/s)	Rate of transitions between grid cells (grid cell size is the same as for spatial density), where a transition is a saccade to another cell (Goldberg and Kotval 1999). A higher rate of transitions indicates less directness.
<b>Duration metrics</b>	
Mean fixation duration (s)	Mean duration of all fixations within a defined period

### The present study

In summary, the question we are trying to answer here is whether eye tracking can be used to determine – within the first few seconds of search – if a person is struggling with clutter on a website, both in the presence and absence of time pressure. This question was then broken down into a set of specific aims: (1) determine the effects of clutter and time pressure on performance, whether searching for words or for images and (2) determine whether eye tracking can be used to trace the effects of clutter during the first few seconds of search. This information would be valuable for interface and user experience designers, providing them with a way to quasi-automatically evaluate their website. This could then pave the way for real-time adjustments if needed. The clutter algorithms of Rosenholtz, Li, and Nakano (2007), which are freely available, were used to classify websites as low or high in clutter as a baseline estimate. These algorithms consist of feature congestion (FC; the difference in colour or luminance features in a display), subband entropy (SE; the amount of redundancy in an image, with more clutter related to less redundancy), and edge density (ED; the frequency of edges).

The eye-tracking metrics that were evaluated in this study are shown in Table 1. These metrics have been proposed for clutter analysis by Moacdieh and Sarter (2015b, 2017a, 2017b). Spread metrics show whether clutter causes dispersion of eye movements across the display, with more spread generally indicating more clutter. Directness measures relate to how ordered and efficient search is in terms of how a user moves towards the target. Previous studies have shown clutter to make search more ordered and systematic (Moacdieh and

Sarter 2015b, 2017a). Finally, duration measures indicate how long a person looked at a particular area and relate clutter primarily to the difficulty extracting information from the display. Longer duration would then indicate more clutter (Beck, Lohrenz, and Trafton 2010), although in other studies it has been found to decrease as people speed up their search under high clutter (Moacdieh and Sarter 2015b). The latter approach is what we expect to occur here, in the absence of any discrimination issues.

In parallel with our specific aims, we articulated the following hypotheses. First (Hypothesis H<sub>1</sub>), we expected that clutter in websites will lead to performance decrements and that time pressure will make search faster but less accurate both for word and image search targets. Second (H<sub>2</sub>), we expected that the eye-tracking metrics of Table 1 would indicate higher spread, more directness, and lower duration in high clutter within the first three seconds of a visual search regardless of time pressure, both for word and image search targets.

## Methods

### Participants

Participants were 50 students (23 women and 27 men) from the American University of Beirut (AUB; average age:  $23.3 \pm 3.4$  years). All participants had self-reported normal or corrected to normal colour vision. The study was approved by the AUB Institutional Review Board.

### Experiment setup

Participants sat at a distance of 45 cm ( $68.2^\circ$  visual angle) from a 24-inch monitor ( $1920 \times 1200$  pixels). A

Tobii X3-120 eye tracker (sampling rate: 120 Hz) was attached to the monitor.

### Stimuli

The research stimuli consisted of 40 colour screenshots of existing e-commerce websites. All of the screenshots, when displayed, filled the whole  $1920 \times 1200$  screen. To include a wide range of websites, the websites were selected to represent seven different categories: general e-commerce websites (e.g. Amazon), clothing, event ticketing, cars, travel, food, and jewellery. In addition, to ensure a wide range of clutter within the websites, four particularly badly cluttered websites were selected among the 40 screenshots (Kenwright 2014). Websites that were relatively sparse or empty were deliberately selected as well. The clutter algorithms of Rosenholtz, Li, and Nakano (2007) were then applied to each screenshot in order to get an initial impression of clutter. The range of FC values across all screenshots was from 2.51 to 12.27 (average of  $5.03 \pm 2.10$ ), the range of SE values was from 1.94 to 4.99 ( $3.47 \pm 0.72$ ), and the range of ED values was from 0.024 to 0.13 ( $0.067 \pm 0.026$ ). This range is wider than the one used by Rosenholtz, Li, and Nakano (2007).

Next, each of the three algorithm values for each screenshot was normalised to obtain a value between 0 and 1. The resulting three normalised values were then averaged to obtain an aggregate clutter score for each screenshot. The screenshots were then ranked in order of clutter based on the aggregate clutter score and divided into 20 low-clutter and 20 high-clutter screenshots. There was a significant difference between the aggregate score of the low and high clutter screenshots ( $t(38) = -7.33, p < .001$ , Cohen's  $d = 2.33$ ), with values of  $0.194 \pm 0.113$  in the low-clutter group and  $0.532 \pm 0.171$  in the high-clutter group.

### Search targets

Half of the screenshots were assigned word search targets and half were assigned image search targets, with both targets selected from the screenshot itself (see Figure 1). The average aggregate clutter score for the screenshots with word targets was  $0.41 \pm 0.27$  and  $0.34 \pm 0.16$  for the screenshots with image targets. There was no significant difference between the aggregate clutter scores in these groups, as examined using a  $t$ -test,  $t(38) = 1.016, p = .316$ .

Eccentricity was the deciding and controlling factor when it came to deciding on targets, both for images

and for words. Targets were selected so as to subtend an eccentricity of around  $15^\circ$  visual angle from the centre of the screen. On average, the eccentricity of the targets was  $15.83 \pm 1.19^\circ$ , with 29 screenshots at  $15^\circ$  exactly. The word targets subtended  $0.91 \pm 0.30^\circ$  vertically and  $5.47 \pm 1.43^\circ$  horizontally. The number of letters in word targets was  $7.20 \pm 2.37$  letters. Note that the desire to control eccentricity as much as possible meant that the size and number of letters for word targets were impossible to control. On the other hand, image targets were all selected to be squares of  $2.54^\circ \times 2.54^\circ$  visual angle.

In order to prevent participants from noticing the fact that all targets were at around the same eccentricity, an additional 10 'dummy screenshots' were selected where the targets had a different eccentricity. Five dummy screenshots had image targets and five had word targets. The data from the dummy screenshots was discarded.

### Experiment design

The independent variables in this study were clutter (low, high) and time pressure (time limit, no time limit), which were both manipulated within-subjects. The low- and high-clutter screenshots represented the two levels of clutter. The time limit simulated the presence of time pressure and was selected to be 10 s, which was the average RT ( $9.85 \pm 13.95$  s) obtained during pilot tests with four participants. The presence of a time limit meant that participants had to complete their search task within this amount of time; otherwise, the trial timed out. In the absence of a time limit, participants had as much time as they needed to complete their tasks. In both cases, participants had the option to give up. Target type (image target or word target) was treated as a blocking variable given that we were not interested in comparing image versus word search; rather, the focus was on clutter and time pressure.

Participants were asked to perform search tasks for the search target (word or image) on each screenshot. Thus each trial consisted of a search task on a given screenshot. In total, each participant had 50 trials to complete: 40 experiment trials and 10 dummy trials (i.e. trials associated with the dummy screenshots). The screenshots were ordered so that participants did 10 sets of five trials each, with each set containing a combination of two low-clutter screenshots (one with an image target and one with a word target) and two high-clutter screenshots (similarly, one image and one word target), as well as a dummy screenshot that

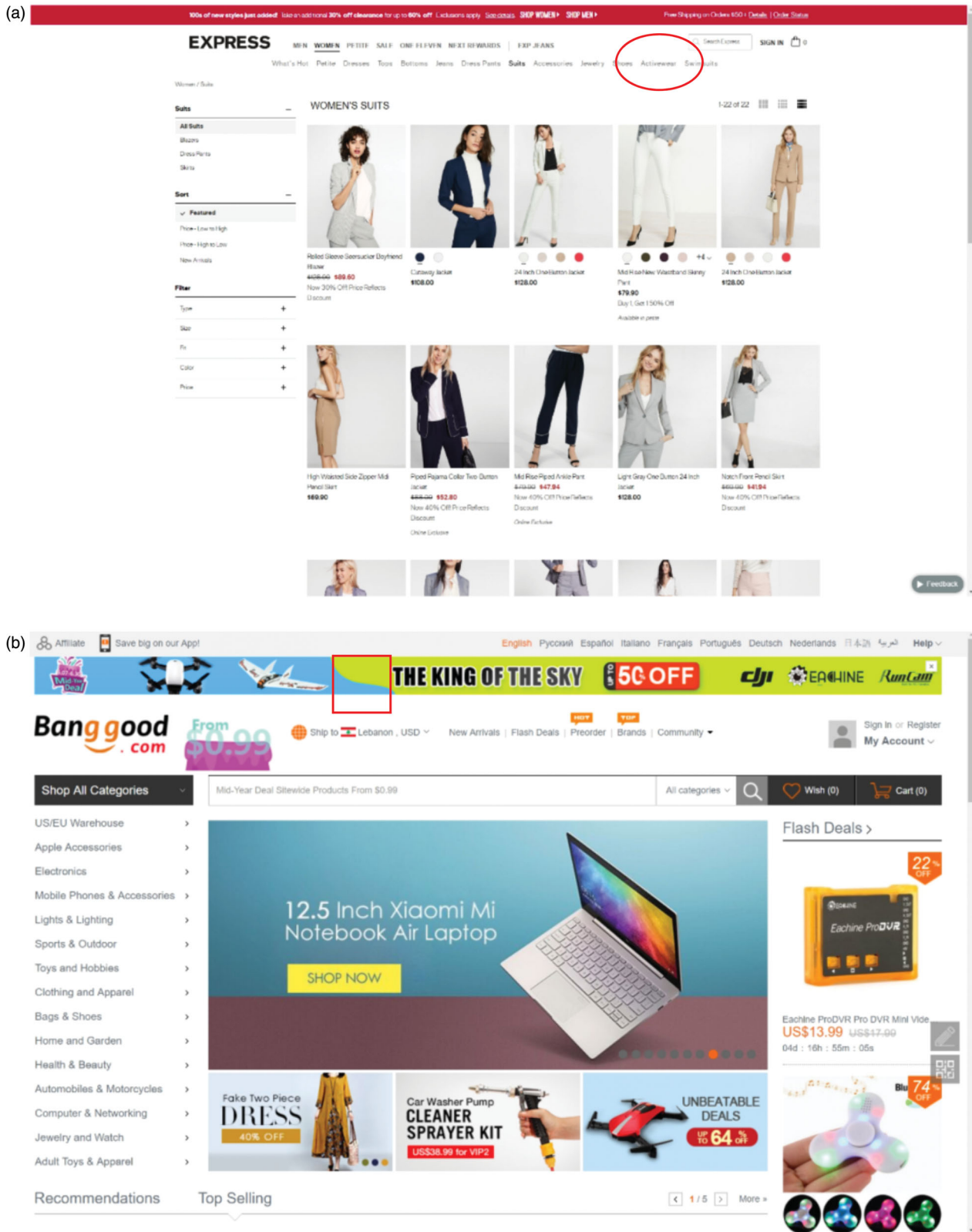



Figure 1. Samples of website screenshots and their corresponding targets (circled in red): (a): word target and screenshot (aggregate clutter score: 0.15); (b) image target and screenshot (aggregate clutter score: 0.35); (c) image target and screenshot (aggregate clutter score: 0.5); (d) word target and screenshot (aggregate clutter score: 1).

(c)


**Wild Discovery**

--EUROPE SEARCH


## EUROPE PACKAGES




**Iconic Aegean: Celestyal Nefeli...**  
Mykonos, Athens, Santorini, Heraklion, Rhodes, Patmos, Kusadasi  
Celestyal Cruises, Sea Cruise, Honeymoon Trains & Cruises  
4 Days , 3 nights




**Celestyal Nefeli: Euphoric Aege...**  
Santorini, Nafplio, Chania, Rhodes, Mykonos, Athens, Izmir, Çeşme  
Celestyal Cruises, Sea Cruise, Honeymoon Trains & Cruises  
7 Days , 6 nights




**Grand Resort Bad Ragaz 5\***  
Zurich  
Spa & Wellness, Honeymoon, Luxury, Honeymoon  
Luxury Hideaways  
8 Days , 7 nights



WILDISCOVERY.COM.LR  
**SiteLock**  
MALWARE-FREE



**SPECIAL OFFER**



(d)

**LINGSCARS.COM**  
Leader of the Pack - Guardian, UK favourite car leasing co from Northeast  
Connect the cars from Ling Valentine, 1 FINECARS in the UK's Specialist Car Leasing website - Go 2018 7 Good 1,887 million in 2017/2018

**I AM LING YOU CAN TRUST ME**

**Menu**

- Home
- Cars
- Customers
- About Ling
- Fun stuff
- Free stuff
- Live staff

**CARS A-Z**

**ABARTH**

**ALFA ROMEO**

**AUDI**



**FANTASTIC front-end JOB**  
media, graphics, movies, music, editing, layout, content, finishing gifts...  
suit a North-East uni student  
**SUITABLE VICTIMS. CLICK HERE**

**Leader of the Pack**  
"Ling Valentine is] Britain's **BIGGEST** individual seller of new cars."  
February 2017 **the guardian**

**MY BEST SELLING CAR LEASING DEALS!**

Model	Price	Leasing Deal
Peugeot 108	£151	3 year cheap car leasing 3-35
Seat Ibiza (to 2016)	£167	3 year cheap car leasing 3-35
Nissan Juke	£177	3 year cheap car leasing 3-35

**Car Leasing Online Service Response Times**

78 customers in 'previous'

Ling median: **0:01:59**

133 customers in 'today'

Ling median: **0:03:36**

No waiting at any time

Figure 1. Continued.

could be any combination of high/low and word/image. There were no target-absent trials.

There were four different, fixed sequences of these 10 sets. Participants were randomly assigned to one of these four sequences. For two of the four sequences, the first 25 trials were all done with a time limit, and then the second set of 25 trials were done with no time limit. The inverse was done for the other two groups. This was done so as to give participants clear instructions at the start of each set of 25 trials, and also in order to not attenuate the effects of time pressure by interspersing no-time limit trials in between. Note that the same screenshots were shown for half of the participants with a time limit and for the other half with no time limit (the time limit was not specific to certain screenshots).

### Dependent variables

The dependent variables were subjective data (for validation), performance data (RT and error rate), and the eye-tracking metrics of Table 1. The subjective data were collected by means of two questionnaires. First, participants filled out a modified NASA Task Load Index (TLX) questionnaire (Hart and Staveland 1988), which asked participants to rate their mental demand, temporal demand, performance, effort, and frustration on a scale from 1 (worst performance or least effort/frustration) to 10 (best performance or most effort/frustration). Second, a post-experiment questionnaire asked participants to rate their impression of clutter in all of the study screenshots on a scale of 1 (least cluttered) to 10 (most cluttered).

RT was calculated from when the screenshot appeared until participants clicked on the location of the target. Any trial that contained an error was not included in the calculation of RT. Three types of errors were considered: miss errors (wrong target), time-out errors (for time limit trials only), and giving-up errors.

For the eye-tracking metrics, the metrics of Table 1 were calculated for the first 3 s of the search. We chose this time window as it was shorter than almost all the response times and had already been tested in a previous study done by Moacdieh and Sarter (2017b), albeit in a different context. The rationale was to start with a time window that has been used with the same eye tracking metrics before, in the hope of decreasing the window in future studies.

### Experiment procedure

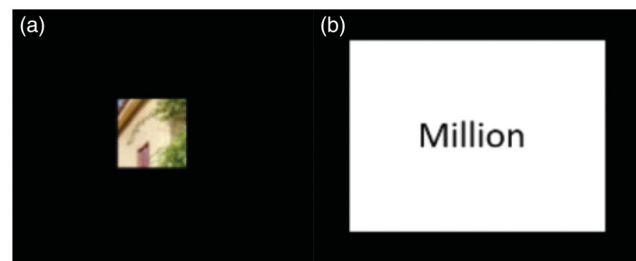
Participants first completed four training tasks. Participants were then told whether they had a time

limit for the next set of 25 trials. Next, the eye tracker was calibrated using a nine-point grid. Then an image or word target would appear (see Figure 2) and participants could look at it as long as they needed. Participants pressed the right arrow key to proceed, at which point a screenshot would appear. Participants had to click on the assumed location of the target with the left mouse key, or right-click anywhere to give up. If the trial timed out, the screenshot would be immediately replaced by the next target. Participants filled out the NASA-TLX questionnaire twice, once after each set of 25 screenshots. There was a 5-min break after the first NASA-TLX questionnaire. The post-experiment questionnaire was filled out at the end of the study. As part of that questionnaire, the participants were shown each of the screenshots that they had seen in the experiment and were asked to give a clutter rating for that screenshot. The screenshots were shown in a randomly assigned order to all participants.

## Results

### Subjective results

The subjective results largely validated our manipulations of clutter and time pressure. The correlation between the subjective ratings of clutter and the algorithm scores were relatively high (greater than 0.7)



**Figure 2.** Screenshots of how the (a) image and (b) word targets would appear to participants. The image target was the same size as the actual target in the image (a square of size  $2.54^\circ \times 2.54^\circ$  visual angle). The word targets were all displayed using the same size black Arial font against a white background. The height of the lowercase letters was around  $1.9^\circ$  visual angle (the width depended on the word), while the size of the white rectangle surrounding the word was always  $10.79^\circ$  (width) by  $8.26^\circ$  (height) visual angle.

**Table 2.** Correlation coefficients between each clutter algorithm and the subjective clutter ratings.

	Correlation coefficients $r_s$ ( $p$ -value)
FC	0.927 ( $p < .001$ )*
SE	0.824 ( $p < .001$ )*
ED	0.711 ( $p < .001$ )*
Aggregate clutter score	0.81 ( $p < .001$ )*

\*Denotes significance at  $p < .05$ .

**Table 3.** Results of the NASA-TLX ratings along the different scales.

NASA TLX Scale (1–10)	Non-timed, median (SD)	Timed cases, median (SD)	Wilcoxon ranked sign test
Mental demand	6 (1.76)	7 (1.91)	$z = -2.932$ ( $p = .03$ )*
Temporal demand	4 (1.99)	8 (1.16)	$z = -6.055$ ( $p < .0005$ )*
Performance	8 (1.42)	6 (1.52)	$z = 4.689$ ( $p < .0005$ )*
Effort	7 (2.02)	8 (1.617)	$z = -3.124$ ( $p = .002$ )*
Frustration	4 (2.74)	7 (2.58)	$z = -3.069$ ( $p = .002$ )*

\*Denotes significance at  $p < .05$ .

and significant for all three algorithms and for the aggregate score (see Table 2). There was also a significant difference between the subjective clutter ratings in the low-clutter ( $4.25 \pm 1.14$ ) and high-clutter ( $7.15 \pm 1.66$ ) conditions ( $t(38) = -6.414$ ,  $p < .001$ ).

In addition, the NASA-TLX results were analysed using a Wilcoxon test since the data significantly deviated from normal based on a Shapiro-Wilk test. There were significant effects of time pressure on all six scales (see Table 3). Participants felt more frustration, mental demand, and temporal demand in the timed cases. They also generally felt they were giving more effort and that their performance was worse.

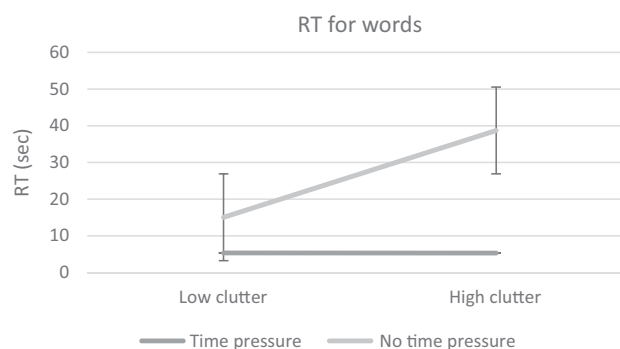
### Performance results

Performance results were analysed using a  $2 \times 2$  repeated-measures analysis of variance (ANOVA). Significance was set at  $p < .05$ , and partial eta-squared ( $\eta_p^2$ ) was used as a measure of effect size. The data were normally distributed, as established using a Shapiro-Wilk test for normality ( $p > .05$ ). Since the independent variables have two levels, the sphericity assumption was met. The ANOVA results are reported for statistically significant results only.

### Word targets

For response time, there was a significant interaction effect between clutter and time pressure  $F(1, 30) = 48.703$ ,  $p < .001$ ,  $\eta_p^2 = .619$  (see Figure 3 and Table 4). There was a simple main effect of clutter in the no time pressure condition  $F(1, 49) = 65.339$ ,  $p < .001$ ,  $\eta_p^2 = .571$ . There was a simple main effect of time pressure in the low-clutter condition,  $F(1, 48) = 63.898$ ,  $p < .001$ ,  $\eta_p^2 = .571$  and also in the high-clutter condition,  $F(1, 30) = 86.665$ ,  $p < .001$ ,  $\eta_p^2 = .743$ . Note that 19 data points were lost due to time out errors in the high-clutter conditions.

For the miss error rate and the giving up error rate, the number of errors overall was so low that no statistical analysis was performed (see Table 4). As for the time-out error rate, a paired-samples t-test revealed a significant difference between the low- and high-clutter conditions,  $t(45) = -7.055$ ,  $p < .005$ .



**Figure 3.** RT for words in each of the four conditions. Error bars represent the standard error of the mean (SEM).

**Table 4.** Summary of the performance results for words.

	Non-timed, mean (SEM)		Timed, mean (SEM)	
	Low clutter	High clutter	Low clutter	High clutter
Response time	15.06 (8.47)	38.72 (21.40)	5.33 (1.64)	5.30 (1.76)
Miss error rate	0.00 (0.00)	0.05 (0.02)	0.00 (0.00)	0.01 (0.01)
Giving up error rate	0.02 (0.01)	0.19 (0.04)	0.02 (0.01)	0.00 (0.00)
Time-out error rate			0.41 (0.04)	0.74 (0.04)

### Image targets

There was no interaction effect between clutter and time pressure. However, there was a main effect of clutter on response time  $F(1, 49) = 9.374$ ,  $p = .004$ ,  $\eta_p^2 = .161$ , and a main effect of time pressure on response time  $F(1, 49) = 18.947$ ,  $p < .001$ ,  $\eta_p^2 = .279$  (see Figure 4 and Table 5).

Similar to the case of words, no statistical analysis was performed on the miss and giving up error rates given the extremely low values overall (see Table 5). No analysis was performed on the time-out error rate for images either for the same reason.

### Eye tracking results

#### Word targets

For words, convex hull area, spatial density, NNI, and transition rate showed the main effects of clutter (see Table 6). There was a significant interaction effect between clutter and time pressure for scan path length per second ( $F(1, 49) = 11.564$ ,  $p = .001$ ,  $\eta_p^2 = .191$ ). However, there was no significant simple main

effect of clutter in either the timed or non-timed conditions. The main effect of clutter for that metric is nevertheless included in Table 6 for completeness.

### Image targets

For images, none of the metrics showed significant interaction effects between clutter and time pressure; however, convex hull area, spatial density, NNI, total fixation number, scan path ratio and backtrack rate showed the main effects of clutter (see Table 7).

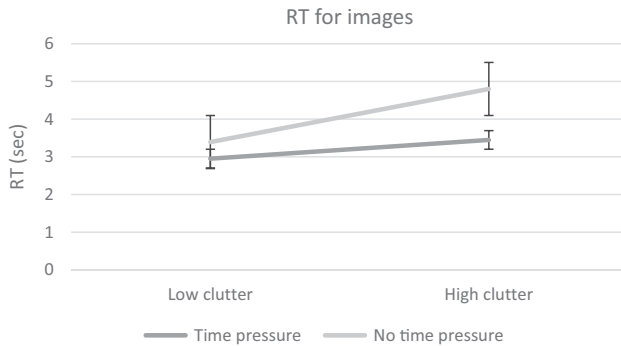


Figure 4. RT for images in all the conditions. Error bars represent SEM.

Table 5. Summary of the performance results for images.

	Non-timed, mean (SEM)		Timed, mean (SEM)	
	Low clutter	High clutter	Low clutter	High clutter
Response time	3.39 (2.09)	4.79 (3.23)	2.95 (1.81)	3.44 (1.18)
Miss error rate	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Giving up error rate	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)
Time-out error rate			0.00 (0.00)	0.04 (0.01)

Table 6. Eye tracking results for words.

Eye tracking metrics	Low clutter (SEM)	High clutter (SEM)	Main effects of clutter
<b>Spread metrics</b>			
Convex hull area	261585.9 (10366.64)	<b>285801.4 (9684.79)</b>	$F(1, 49) = 6.668$ , $p = .013$ , $\eta_p^2 = .120$
Spatial density	0.025 (0.0005)	<b>0.027 (0.0005)</b>	$F(1, 49) = 11.982$ , $p = .001$ , $\eta_p^2 = .196$
Nearest neighbour index	0.529 (0.01)	<b>0.553 (0.01)</b>	$F(1, 49) = 4.652$ , $p = .036$ , $\eta_p^2 = .087$
Total fixation number	9.042 (0.22)	<b>9.448 (0.20)</b>	$F(1, 49) = 6.617$ , $p = .013$ , $\eta_p^2 = .119$
<b>Directness metrics</b>			
Scanpath length per second	0.71 (0.018)	0.70 (0.015)	$F(1, 49) = 0.696$ , $p = .408$ , $\eta_p^2 = .014$
Mean saccade length	<b>252.9 (5.69)</b>	242.2 (6.59)	$F(1, 49) = 9.6$ , $p = 0.003$ , $\eta_p^2 = .164$
Backtrack rate	0.0013 ( $3.7 \times 10^{-5}$ )	0.0012 ( $3.6 \times 10^{-5}$ )	$F(1, 49) = 1.151$ , $p = .259$ , $\eta_p^2 = .023$
Transition rate	0.0025 ( $4.8 \times 10^{-5}$ )	<b>0.0026 (<math>4.7 \times 10^{-5}</math>)</b>	$F(1, 49) = 0.308$ , $p = .578$ , $\eta_p^2 = .009$
<b>Duration metric</b>			
Mean fixation duration	<b>198.8 (3.43)</b>	192.5 (3.54)	$F(1, 49) = 6.336$ , $p = .015$ , $\eta_p^2 = .114$

Bold entries indicate significantly larger values.

## Discussion

In general, the subjective clutter ratings and NASA-TLX results validated the manipulation of clutter and time pressure, respectively. We can thus use the results to examine our two hypotheses.

### H<sub>1</sub>: Effects of clutter and time pressure on performance

As expected, the presence of clutter led to an increase in RT, in line with other studies on clutter (Beck et al. 2012; Bravo and Farid 2008; Henderson, Chanceaux, and Smith 2009; Moacdieh and Sarter 2015b; Neider and Zelinsky 2011). This was the case in this study both for word and for image targets. In the case of word targets, there was an interaction effect between clutter and time pressure, but this was due to the fact that participants were largely unable to answer in time during the time pressure conditions. This was seen in the very large time-out errors that occurred – otherwise, the error rates were negligible. For the image targets, given that search was easier (as evidenced by the lower RTs), the expected effects were observed in full: higher RTs due to clutter and lower RTs due to time pressure. We had also expected that there would be larger error rates in the presence of time pressure; however, in general, the error rates were largely negligible.

These results are in contrast to those found in Moacdieh and Sarter (2015b), where time pressure worsened the effects of clutter in the context of search in EMRs, both in terms of longer RT and higher error rate. The difference could be due to the fact that Moacdieh and Sarter (2015b) manipulated stress; this

**Table 7.** Eye tracking results for images.

Eye tracking metrics	Low clutter (SEM)	High clutter (SEM)	Main effects of Clutter
<b>Spread metrics</b>			
Convex hull area	159920.0 (6748.56)	<b>203939.4 (10603.17)</b>	$F(1, 49) = 5.742$ , $p = .020$ , $\eta_p^2 = .105$
Spatial density	0.018 (0.0004)	<b>0.02 (0.0004)</b>	$F(1, 49) = 16.365$ , $p < .001$ , $\eta_p^2 = .250$
Nearest neighbour index	0.47 (0.0102)	<b>0.51 (0.0121)</b>	$F(1, 49) = 6.250$ , $p = .016$ , $\eta_p^2 = .113$
Total fixation number	6.45 (0.17)	<b>7.21 (0.16)</b>	$F(1, 49) = 20.859$ , $p < .001$ , $\eta_p^2 = .424$
<b>Directness metrics</b>			
Scanpath length per second	0.776 (0.018)	0.814 (0.02)	$F(1, 49) = 0.003$ , $p = .959$ , $\eta_p^2 = .000$
Mean saccade length	264.3 (5.10)	277.3 (7.47)	$F(1, 49) = 0.458$ , $p = 0.502$ , $\eta_p^2 = .009$
Backtrack rate	0.0011 ( $3.73 \times 10^{-5}$ )	<b>0.0012 (<math>4.09 \times 10^{-5}</math>)</b>	$F(1, 49) = 5.022$ , $p = .030$ , $\eta_p^2 = .093$
Transition rate	0.00224 ( $5.65 \times 10^{-5}$ )	0.00227 ( $4.97 \times 10^{-5}$ )	$F(1, 49) = 0.303$ , $p = .585$ , $\eta_p^2 = .006$
<b>Duration metric</b>			
Mean fixation duration	231.512 (4.35)	223.75 (3.85)	$F(1, 49) = 1.536$ , $p = .221$ , $\eta_p^2 = .030$

Bold entries indicate significantly larger values.

**Table 8.** summary of the eye tracking results in comparison to selected previous studies that analysed the whole search period.

	Previous literature*	Present study (words)	Present study (images)
<b>Spread metrics</b>			
Convex hull area <sup>a,b</sup>	↑	↑	↑
Spatial Density <sup>a,b</sup>	↑	↑	↑
Nearest neighbour index <sup>a,b</sup>	↑	↑	↑
Total fixation number <sup>a,c,d,e</sup>	↑	↑	↑
<b>Directness metrics</b>			
Scanpath length per second <sup>b</sup>	↓		
Mean saccade length <sup>a,b,d</sup>	↓	↓	
Backtrack rate <sup>a</sup>	↓		↑
Transition rate <sup>b,d</sup>	↓	↑	
<b>Duration metric</b>			
Mean fixation duration <sup>a,b,d</sup>	↑↓	↓	-

Note: Arrows indicate what happened to that metric with an increase in clutter.

\*Studies included: <sup>a</sup>Moacdieh and Sarter (2017a); <sup>b</sup>Moacdieh and Sarter (2015b); <sup>c</sup>Yoon, Lim, and Ji (2015); <sup>d</sup>Beck et al. (2012);

<sup>e</sup>Beck, Lohrenz, and Trafton (2010).

included time pressure as a major component, but also involved making the medical tasks more critical and giving participants incentives. This approach to manipulating stress – in terms of time pressure and incentives – was also adopted by Moacdieh and Sarter (2017a), but in that case, the results were more consistent with the present study: high clutter caused a decrease in time pressure and no effect on the error rate. The similarity between the present study and Moacdieh and Sarter (2017a), which involved the search for icons in a graphics display, is the simplicity of the search tasks, even though the current study was in a more realistic environment. Moacdieh and Sarter (2015b) also demonstrated that increased task difficulty accentuates the effect of clutter, where a difficult task was defined there as the presence of more

than one search target and the need for comparing the two before coming to a decision. It could be that in the case of simple search detection, time pressure does not have an effect, whereas it does have more of an effect at higher levels of cognitive processing.

### ***H<sub>2</sub>: Eye-tracking metrics during the first three seconds***

Having established that performance decrements occurred due to clutter, the follow-up and more crucial question was whether these performance effects would translate into differences in eye-tracking metrics in the early seconds of search. This appeared to be the case, with several eye-tracking metrics showing significant main effects of clutter, even within the first

three seconds. This was evident both for images and for words, and with mostly no significant interaction effects with time pressure. This suggests that a practitioner could conceivably be able to use these metrics reliably in a variety of conditions. The only metric to show an interaction effect with time pressure was scan path length per second, in which case there was no significant simple main effect of clutter in either of the time pressure conditions.

In general, the spread metrics were the ones that appeared to be most promising during the first three seconds, with all four metrics increasing with increasing clutter for both word and image targets. This is consistent with our hypothesis and with prior research on clutter over the whole search period (e.g. Moacdieh and Sarter 2017a; Yoon, Lim, and Ji 2015), suggesting that even as early as just the first 3 s, people's spread of attention was noticeably larger in the case of high clutter. Among the directness metrics, only mean saccade length and transition rate (for words) and backtrack rate (for images) were different in the case of high clutter. However, there was some inconsistency between our results and previous research. Mean saccade length decreased with increasing clutter, which is consistent with our hypothesis about more directness under high clutter. On the other hand, both transition rate and backtrack rate increased instead. The effect of clutter on directness metrics is then not as clear as for spread metrics. It would seem that when searching for words under high clutter, participants in the first 3 s transitioned to many different areas but with short saccades, whereas with images they tended to go back and forth a lot more across the display. This could be because the image targets were more salient, whereas with words one had to move more carefully and deliberately through the display. Finally, mean fixation duration was not affected by clutter, suggesting that, even though fixation duration is a popular metric to use in clutter research (e.g. Beck et al. 2012), it does not always reflect the changes between low and high clutter, as was also the case in Moacdieh and Sarter (2017a). This may be because mean fixation duration is affected by various factors, such as confusion and interest. It could also be that mean fixation duration increases later on during search, as users resort to a more deliberate approach as time goes by (Beck et al. 2012). The results provide a warning about using mean fixation duration to study clutter early on during the search process. Table 8 summarises the eye-tracking findings in comparison to previous findings over the whole search period.

As for the few studies that have tried to look at the effects of clutter early on during search, a previous study done by Moacdieh and Sarter (2015b) on EMRs did not show any significant effects of clutter in the first 4 s of search on the same eye-tracking metrics. Another study done by the same authors (Moacdieh and Sarter 2017b) built a predictive model of RT over the first 3 s based on three different metrics: mean saccade length, scan path length, and mean fixation duration. However, it was not clear what association each of them had with increased clutter. It would appear that the current study has taken a significant step in showing that people react differently to cluttered versus uncluttered displays and that eye-tracking metrics – and, in particular, spread metrics – can be used to detect that same difference over just the first 3 s of search.

## Conclusion

In summary, this study provided answers to the questions of whether and how eye tracking metrics can be used to trace people's reactions to cluttered displays. In response to *whether* this is possible, several eye-tracking metrics were significantly different when first viewing a cluttered display as compared to a non-cluttered one. This held even if the user is under time pressure, and whether they are searching for images or words. In a departure from many previous studies, the types of displays used were existing websites, not specialised or non-realistic displays. As for *how* this can be done, results showed that spread metrics are very useful in this regard. These results provide support for the creation of real-time, adaptive websites where these metrics are calculated in real-time at the start of the search and used to determine whether the user finds the display cluttered or not. Adjustments can then be made, such as fading out some of the less important information or preventing additional pop-ups from further disturbing the user. Assuming that eye tracking is likely to become widespread in personal computers and laptops, such an ability to assess a person's reactions can give a designer an edge and an endless ability to customise.

Further research will look into tackling the main limitation of this study, which is that the size and salience of targets were not controlled (since real-life websites were used). The next step would be to create realistic websites where all factors can be perfectly controlled. This would then provide further insight into which eye-tracking metrics are best to use. Another limitation was that the cut-off time limit set

made the time-out errors very high in the case of words, so future studies could relax the time limit and further explore the effects of time pressure in different forms. Regression analysis using the clutter scores could also be done instead of the analysis of variance adopted here. Also, the differences between image and word search could be further explored, and the interaction between clutter and the type of search target could be analysed in detail. Finally, another long-term initiative would be to try to decrease the 3-s window even further in order to potentially detect people's reactions to clutter within milliseconds. The 3-s time window used here was only a starting point, and real-time display adjustments would need an even smaller window if they are to be effective. Doing that would further refine which eye-tracking metrics are best to use.

## Acknowledgement

The authors would like to thank Hussein Jundi for his help in running this study and writing the Matlab code.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

Nadine Marie Moacdieh  <http://orcid.org/0000-0002-7677-7946>

## References

- Asher, M. F., D. J. Tolhurst, T. Troscianko, and I. D. Gilchrist. 2013. "Regional Effects of Clutter on Human Target Detection Performance." *Journal of Vision* 13 (5): 25. doi:10.1167/13.5.25.
- Beck, M. R., M. C. Lohrenz, and J. G. Trafton. 2010. "Measuring Search Efficiency in Complex Visual Search Tasks: Global and Local Clutter." *Journal of Experimental Psychology. Applied* 16 (3): 238–250. doi:10.1037/a0019633.
- Beck, M. R., M. Trenchard, A. van Lamsweerde, R. R. Goldstein, and M. Lohrenz. 2012. "Searching in Clutter: Visual Attention Strategies of Expert Pilots." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1411–1415. Los Angeles, CA: Sage Publications.
- Bravo, M. J., and H. Farid. 2008. "A Scale Invariant Measure of Clutter." *Journal of Vision* 8 (1): 23–29. doi:10.1167/8.1.23.
- Di Nocera, F., M. Camilli, and M. Terenzi. 2007. "A Random Glance at the Flight Deck: Pilots' Scanning Strategies and the Real-Time Assessment of Mental Workload." *Journal of Cognitive Engineering and Decision Making* 1 (3): 271–285. doi:10.1518/155534307X255627.
- Ehinger, K. A., and R. Rosenholtz. 2016. "A General account of Peripheral Encoding Also Predicts Scene Perception Performance." *Journal of Vision* 16 (2): 13. doi:10.1167/16.2.13.
- Fabrikant, S. I., S. R. Hespanha, and M. Hegarty. 2010. "Cognitively Inspired and Perceptually Salient Graphic Displays for Efficient Spatial Inference Making." *Annals of the Association of American Geographers* 100 (1): 13–29. doi:10.1080/00045600903362378.
- Goldberg, J. H., and X. P. Kotval. 1999. "Computer Interface Evaluation Using Eye Movements: Methods and Constructs." *International Journal of Industrial Ergonomics* 24 (6): 631–645. doi:10.1016/S0169-8141(98)00068-7.
- Hart, S. G., and L. E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research." In *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland.
- Hegarty, M., K. De Leeuw, and B. Bonura. 2008. "What Do Spatial Ability Tests Really Measure." Paper presented at the 49th Meeting of the Psychonomic Society, Chicago, IL, November 13–16.
- Henderson, J., M. Chanceaux, and T. Smith. 2009. "The Influence of Clutter on Real-World Scene Search: Evidence from Search Efficiency and Eye Movements." *Journal of Vision* 9 (1): 32. doi:10.1167/9.1.32.
- Jahanian, A., S. Keshvari, and R. Rosenholtz. 2018. "Web Pages: What Can You See in a Single Fixation?" *Cognitive Research* 3 (1): 14.
- Kaber, D. B., J. M. Riley, K. W. Tan, and M. R. Endsley. 2001. "On the Design of Adaptive Automation for Complex Systems." *International Journal of Cognitive Ergonomics* 5 (1): 37–57. doi:10.1207/S15327566IJCE0501\_3.
- Kaufmann, K., and D. B. Kaber. 2010. "The Influence of Individual Differences in Perceptual Performance on Pilot Perceptions of Head-up Display Clutter." In *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*, 70–74. Santa Monica, CA: Human Factors and Ergonomics Society.
- Kenwright, S. 2014. "Top 10 Worst Websites You'll Wish You Hadn't Seen." *The Edit*, July 30. <https://edit.co.uk/blog/top-10-worst-websites/>
- Moacdieh, N. M., and N. Sarter. 2015a. "Display Clutter: A Review of Definitions and Measurement Techniques." *Human Factors* 57 (1): 61–100. doi:10.1177/0018720814541145.
- Moacdieh, N. M., and N. Sarter. 2015b. "Clutter in Electronic Medical Records: Examining Its Performance and Attentional Costs Using Eye Tracking." *Human Factors* 57 (4): 591–606. doi:10.1177/0018720814564594.
- Moacdieh, N. M., and N. Sarter. 2017a. "The Effects of Data Density, Display Organization, and Stress on Search Performance: An Eye Tracking Study of Clutter." *IEEE Transactions on Human-Machine Systems* 47 (6): 886–895. doi:10.1109/THMS.2017.2717899.
- Moacdieh, N. M., and N. Sarter. 2017b. "Using Eye Tracking to Detect the Effects of Clutter on Visual Search in Real Time." *IEEE Transactions on Human-Machine Systems* 47 (6): 896–902. doi:10.1109/THMS.2017.2706666.
- Naylor, J. 2010. "The Influence of Dynamics, Flight Domain and Individual Flight Training & Experience on Pilot Perception of Clutter in Aviation Displays." Master's thesis, North Carolina State University.

- Neider, M. B., and G. J. Zelinsky. 2011. "Cutting through the Clutter: Searching for Targets in Evolving Complex Scenes." *Journal of Vision* 11 (14): 7. doi:[10.1167/11.14.7](https://doi.org/10.1167/11.14.7).
- Paivio, A., and I. Begg. 1974. "Pictures and Words in Visual Search." *Memory & Cognition* 2 (3): 515–521. doi:[10.3758/BF03196914](https://doi.org/10.3758/BF03196914).
- Pankok, C., Jr., and D. Kaber. 2018. "The Effect of Navigation Display Clutter on Performance and Attention Allocation in Presentation- and Simulator-Based Driving Experiments." *Applied Ergonomics* 69: 136–145. doi:[10.1016/j.apergo.2018.01.008](https://doi.org/10.1016/j.apergo.2018.01.008).
- Pieters, R., and L. Warlop. 1999. "Visual Attention during Brand Choice: The Impact of Time Pressure and Task Motivation." *International Journal of Research in Marketing* 16 (1): 1–16. doi:[10.1016/S0167-8116\(98\)00022-6](https://doi.org/10.1016/S0167-8116(98)00022-6).
- Rosenholtz, R., Y. Li, and L. Nakano. 2007. "Measuring Visual Clutter." *Journal of Vision* 7 (2): 17–22. doi:[10.1167/7.2.17](https://doi.org/10.1167/7.2.17).
- Schons, V. W., and C. D. Wickens. 1993. *Visual Separation and Information Access in Aircraft Display Layout (Technical Report ARL-93-7/NASA-A 3 I-93-1)*. Savoy, IL: University of Illinois, Aviation Research Laboratory.
- Tuch, A. N., J. A.argas-Avila, K. Opwis, and F. H. Wilhelm. 2009. "Visual Complexity of Websites: Effects on Users' Experience, Physiology, Performance, and Memory." *International Journal of Human-Computer Studies* 67 (9): 703–715. doi:[10.1016/j.ijhcs.2009.04.002](https://doi.org/10.1016/j.ijhcs.2009.04.002).
- Wickens, C. D., A. Nunes, A. L. Alexander, and K. Steelman. 2005. *3D Navigation and Integrated Hazard Display in Advanced Avionics: Workload, Performance, and Situation Awareness Display Clutter Issues in SVS Design (AHFD-05-19/NASA-05)*. Urbana, IL: University of Illinois Human Factors Division.
- Wolfe, J. M., and T. S. Horowitz. 2017. "Five Factors That Guide Attention in Visual Search." *Nature Human Behaviour* 1 (3): 0058. doi:[10.1038/s41562-017-0058](https://doi.org/10.1038/s41562-017-0058).
- Wolfe, J. M., M. L. H. Vö, K. K. Evans, and M. R. Greene. 2011. "Visual Search in Scenes Involves Selective and Nonselective Pathways." *Trends in Cognitive Sciences* 15 (2): 77–84. doi:[10.1016/j.tics.2010.12.001](https://doi.org/10.1016/j.tics.2010.12.001).
- Yoon, S. H., J. H. Lim, and Y. G. Ji. 2015. "Perceived Visual Complexity and Visual Search Performance of Automotive Instrument Cluster: A Quantitative Measurement Study." *International Journal of Human-Computer Interaction* 31 (12): 890–900. doi:[10.1080/10447318.2015.1069661](https://doi.org/10.1080/10447318.2015.1069661).