Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-024-02938-y

In simple but challenging search tasks, most errors are stochastic

Jeremy M. Wolfe[1,2] · **Johan Hulleman**[3] · **Ava Mitra**[1] · **Wentao Si**[4]

Accepted: 7 July 2024

Abstract
In visual search tasks in the lab and in the real world, people routinely miss targets that are clearly visible: so-called look but fail to see (LBFTS) errors. If search displays are shown to the same observer twice, we can ask about the probability of joint errors, where the target is missed both times. If errors are "deterministic," then the probability of a second error on the same display–given that the target was missed the first time–should be high. If errors are "stochastic," the probability of joint errors should be the product of the error rate for first and second appearances. Here, we report on two versions of a *T* among *L*s search with somewhat degraded letters to make search more difficult. In Experiment 1, *T*s could either appear amidst crowded "clumps" of *L*s or more in isolation. Observers made more errors when the *T* was in a clump, but these errors were mainly stochastic. In Experiment 2, the task was made harder by making *T*s and *L*s more similar. Again, errors were predominantly stochastic. If other, socially important errors are also stochastic, this would suggest that "double reading," where two observers (human or otherwise) look at each stimulus, could reduce overall error rates.

Keywords Visual attention · Visual search · Error · Stochastic and deterministic errors

Introduction

In visual search tasks in the laboratory and in the broader world, people make mistakes. Someone is looking for a target amidst nontarget stimuli, and they either fail to find it (a "miss" or "false negative") or they declare it to be present when it is not (a "false alarm" or "false positive"). The study of errors in search is an old one (Chun & Wolfe, 1996; Nartker et al., 2023; Ruckmick, 1926; Titchener, 1924; Zenger & Fahle, 1997) because it is an important problem. As a society, we have designed tasks from cancer screening (e.g., Kundel, 2007), to airport security (e.g., Meuter & Lacherez, 2016), to lifeguarding at the neighborhood pool (Sharpe et al., 2023) where errors, especially false-negative errors, are very costly. Even under less dire circumstances, errors are likely to be undesirable. We do not want to miss typos in our correspondence. We do not want to walk right past our car in the parking lot. Thus, it is in our interest to understand and, if possible, reduce these errors.

Errors do not arise from a single cause. Modulations of the state of the searcher are important. Unsurprisingly, fatigued searchers make more errors (Hanna et al., 2018; Krupinski, 2010; Meuter & Lacherez, 2016). Even when they are not tired, individuals experience a "vigilance decrement" in performance if forced to perform a task for an extended period of time (Davies & Parasuraman, 1982; Greenlee et al., 2022; Rubinstein, 2020; Thomson et al., 2015). Observers who are responding too quickly will make more errors, too: the classic speed–accuracy trade-off (Heitz, 2014). Looking more closely at the search process itself, eye tracking can be used to distinguish between three broad classes of false-negative errors (Kundel et al., 1978): search errors, where the eyes never fixate on or near the target; recognition errors, where

Significance People miss visual stimuli. From typos in manuscripts to tumors in x-rays, people miss stimuli that they want to find, even when the target is known and even though the relevant stimulus may be, literally, right in front of their eyes. Understanding the causes of these 'look but fail to see' errors is important if we want to reduce their frequency. Here, we use a novel technique to show that the bulk of the errors in a basic visual search task appear to be "stochastic," occurring randomly rather than being determined by specific locations of targets or configurations of stimuli.

✉ Jeremy M. Wolfe
jwolfe@bwh.harvard.edu

[1] Brigham and Women's Hospital, 900 Commonwealth Ave, 3rd Floor, Boston, MA 02215, USA

[2] Harvard Medical School, Boston, MA, USA

[3] University of Manchester, Manchester, UK

[4] Dartmouth College, Hanover, NH, USA

Published online: 19 August 2024

the eyes do fixate on or near the target but only for a fraction of a second; and decision errors, where the observer spends time looking at a target but fails to correctly classify it (Waite et al., 2016; Wolfe, Lyu et al., 2022b).

Search and recognition errors can be lumped together as "perceptual" errors (Bruno et al., 2024). These are errors where the target can be readily identified when they are pointed out after it has been missed. Decision errors are different. They might be a matter of inadequate expertise (Oh, is that what a kumquat looks like?) or the stimulus might be fundamentally ambiguous; for instance, a clearly visible item that might or might not be a tumor. Radiologists' perceptual errors can trigger malpractice suits (Berlin & Hendrix, 1998) when a finding is "retrospectively visible" (Kouskos et al., 2004). In the driving literature, these errors (especially recognition errors) are known as "look but fail to see" (LBFTS) errors (Hills, 1980), building on the common statements by drivers asserting that they had looked, let us say, to the left but did not see the other vehicle that they subsequently hit.

There are multiple contributors to LBFTS errors (Wolfe, Kosovicheva et al., 2022a). Individuals do not fully process everything in the useful field of view ([UFOV], or functional visual field [FVF]) surrounding the current point of fixation. This might occur because they only select a subset of items in the FVF for adequate processing (Wu & Wolfe, 2022), or perhaps because summary statistics computed across a functional visual field that does contain a target may not always allow for that target's detection (Rosenholtz et al., 2012). In selecting items to process, observers can be "guided" to likely candidates if those candidates have the right features or are in the right locations (Wolfe, 2021). However, under some circumstances, they can be "misguided" away from the target. In the classic Simons and Chabris (1999) inattentional blindness (gorilla) experiment, observers were more likely to miss the gorilla if they were attending to actors in white shirts than when they attended to those in black shirts. Attending to white-shirted players guided observers away from a black gorilla. Observers may also fail to register a fixated and/or attended item if they end their processing of that item or set of items before identification is complete. If object identification is understood as a random-walk accumulation of information toward an identification boundary (Schall, 2019), one can imagine that, in some cases, the accumulation is slow. At some point, an observer must conclude (probably unconsciously/implicitly) that it is time to move on without obtaining a clear answer. In some cases, this might involve abandoning an identifiable target before it was identified (Wolfe, Kosovicheva et al., 2022a).

There are other error mechanisms; for example, satisfaction of search (Berbaum et al., 2019), also known as "subsequent search misses" (Cain et al., 2013). Whatever the cause of a miss error, it is worth asking why *this particular target* was missed. We can identify two possibilities. The error could be *determined* by the stimulus. To give a trivial example, an item might not be seen because it was hidden or invisible. Less trivially, a clearly visible item might be missed regularly because it is presented in the wrong place or at the wrong time. Alternatively, an error could be random or *stochastic*; a situation where one item in a display is no more likely to be missed than another even though some percentage of items is missed by chance. Again, to offer a trivial example: If a ring of 12 items is presented around fixation for only 100 msec, observers will miss targets at a high rate. They will simply not have the time needed to process all 12 items. If the items are equally detectable and the observer does not have some positional bias, those errors will be stochastic. On the other hand, if the observer reliably starts at the top of the ring and processes items in a clockwise direction, the errors at 10 or 11 o'clock could be considered deterministic.

Li et al. (2024) introduced a method to distinguish whether errors in a task were stochastic or deterministic. Each display in a simple search for a *T* among *L*s was presented twice. If the cause of an error was entirely deterministic, then the probability of missing the item the second time, given that it was missed the first time, would be 1.0. If the errors were stochastic, then the probability of missing both instance #1 and instance #2 would be the probability of missing instance #1 multiplied by the probability of missing the target on instance #2. They would be statistically independent. There are some possible complications, however. For example, *O*s could guess, correctly or otherwise. They could also learn during the task, reducing the error rates on later appearances of the target. Regardless, in most circumstances, there would remain a clear difference between the predictions of deterministic and stochastic accounts. The data could lie between these extreme possibilities if errors were a mix of both types of error.

In the first experiment of Li et al. (2024), the target was a white letter *T* among *L*s in various orientations on a uniform gray background. In this condition, errors were quite clearly stochastic in nature. The chance of missing both instance 1 and 2 of a display was close to P(miss1) × P(miss2). In the second experiment, targets of different contrasts were presented on a variegated background of $1/f^{1.3}$ noise (which roughly mimics the texture of a breast x-ray). In this case, the joint errors fell in between the predictions of a stochastic and a deterministic account. When the results were examined as a function of the contrast of the target, it was clear that the deterministic component of the results was driven by the low contrast targets. That is, somewhat unsurprisingly, if a target was hard to see, it was more likely to be missed on both of its appearances. It is important to note that these twice-missed targets were still visible. As discussed above, it is not interesting to discover that invisible targets are missed.

The most interesting LBFTS errors are those where the target is not difficult to see but is missed, nevertheless. In the present paper, we used stimuli that were of high contrast in all cases but rendered potentially more difficult to find

either by the effects of crowding (Rosenholtz et al., 2012; Zhang et al., 2015) or by using more difficult but still high contrast "offset" *T* and *L* stimuli (see Figs. 1, 2, 3, 4, 5, 6 and 7). To anticipate our findings, both of these searches produced sizable error rates, and, in both cases, the errors appeared to be primarily, but not entirely, stochastic.

Experiment 1: Manipulating task difficulty using stimulus crowding

Participants

Li et al. (2024) aimed for 20 participants. We added a crowding factor to the design of the experiment (described below), so we doubled the planned number to 40. In practice, we recruited 38 participants (15 males, 23 females, $M = 30.10$, $SD = 8.25$, min = 18, max = 50) through Prolific, an online crowdsourcing platform. All participants reported normal or corrected-to-normal vision and accepted an informed consent form presented before they began the experiment. Participants received 9 U.S. dollars in compensation for their participation in the experiment, which took an average of 55 minutes to complete (min = 33, max = 84).

Stimuli and apparatus

In Experiment 1, observers searched for a *T* among *L*s, as in the Li et al. (2024) paper. The vertical and horizontal line segments of the letters were 0.034 × screen height. (Stimuli are described in screen height units because the experiments were conducted online. As a result, we could only control relative size and position of stimuli). Letters could be randomly oriented in 30 deg steps from 0 to 330 deg. The entire display was a square, 0.7 screen height units on a side, centered in the middle of the screen. Items could be scattered pseudorandomly or grouped into a crowded cluster of four items. When present, the target *T* had a 50% chance of appearing as a member of a cluster among clustered and independent distractors. For the other 50% of the trials, the target appeared independently, spatially separated from other items in the display. To create the displays, the screen was first divided into a regular 16 × 16 grid. A cluster was created from four adjacent locations in this grid. Each cluster was 0.25 screen height away from any other. Two clusters were generated when set size was 18, while four clusters were generated when set size was 36. Each item within a cluster was then pseudo-randomly wiggled vertically and horizontally by 0.005 screen height from its original location. As a result, letters within a cluster were spaced from .062 to .088 screen height units apart from each other. Each un-clustered item was placed randomly, under the constraint that it was 0.125 of screen height away from any other items. The distribution of clustered and un-clustered targets was essentially the same across the experiment. Letters were white on a gray background. Again, due to the nature of online testing, the precise contrast levels cannot be specified. In order to make the task more difficult and, thus, more likely to produce errors, *T*s and *L*s were drawn with their two line segments slightly intersecting. Instead of *T* or *L* line junctions, all line crossings were +s. An example of a target-present trial is shown in Fig. 1. Note that, once found, it is clear that the *T* is a *T*.

Design and procedure

Participants were instructed to search for the letter *T* among *L*s. They used the "j" key to indicate that they had found the target and the "f" key to report that they had not. The stimuli were present until response. If the space bar was pressed within 1 second after the initial response, the response would be reversed, allowing observers to correct motor errors. Reversed trials were excluded from analysis along with their paired trials. Targets were present on 50% of trials. Feedback was provided only after every block of 100 trials, where the percentage correct was displayed for that block. Set sizes were 18 and 36, crossed with target presence/absence. As noted, 50% of targets were presented in crowded clusters. There were 25 trials for each of the eight conditions (two set sizes × target presence/absence × clustered/un-clustered). The clustered/un-clustered variable had no impact on the appearance of target absent trials. Each of these 200 trials was presented twice. The resulting 400 total trials were shuffled randomly so the repetitions of a trial could be separated from 1 to 399 positions. Participants went through an eight-trial practice session illustrating all possible combinations of conditions before they started the experiment. We did not have extensive practice or performance criteria to enter the experiment since we were interested in errors.

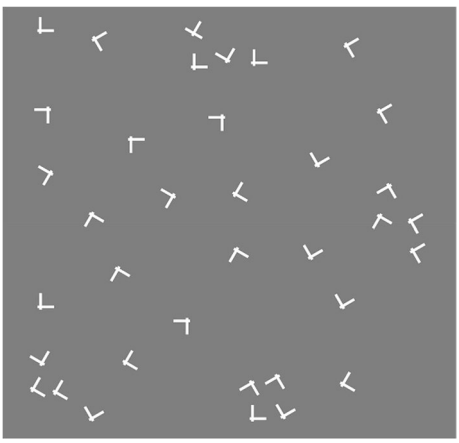

Fig. 1 Sample target-present trial for Experiment 1

Data exclusion

Participants were removed if they made more than 5% motor errors or if their d' was less than 1.0. This removed four observers. For the remaining 34 observers, we analyzed the reaction times (RTs) in each observer × set size × target presence/absence condition. In each cell, we removed any trial with an RT greater than the cell mean plus 2.5 standard deviations for that cell. We removed trials with RTs less than 200 msec as errors of anticipation. We also removed any trial with a motor error. If a trial was removed from analysis, its partner trial was also removed. This cleaning left 98.8% of trials for the 34 observers.

Results

Figure 2 shows the average RT for each observer as a function of Set Size for the first and second appearance of a trial and for crowded and uncrowded targets.

For the present trials, there is an obvious set size effect (three-way analysis of variance [ANOVA]), $F(1, 66) = 67.2$, $p < .0001$, partial eta square = 0.50. There was also an effect of first versus second appearance, with the second appearance RTs being ~300-msec faster than the first, $F(1, 66) = 35.6$, $p < .001$, partial eta-square = 0.35. This can be seen as a general learning effect since the second instances of each display, of course, appear later in the sequence of trials. It is possible that there is a more trial-specific effect, akin to repetition priming or contextual cueing ("Oh," a hypothetical Observer says, implicitly, "the last time I saw this pattern, the target was in that location"; Chun & Jiang, 1998). However, contextual cueing typically involves multiple presentations of the same display and here each display was only presented twice. Priming would be more likely with immediate repetition of the display, and that happened very rarely in this experiment. There was no effect of crowding on RT, $F(1, 66) = 0.11$, $p = .74$, partial eta-square = 0.0017. No interactions approached significance (all ps > .15). The message in the RT data is that it does not take longer to find or to process crowded stimuli. For the absent trials, there is a clear effect of set size, $F(1, 33) = 88.90$, $p < .0001$, no effect of first versus second appearance, and no interaction (all ps > .30).

The false negative/miss error data are shown in Fig. 3.

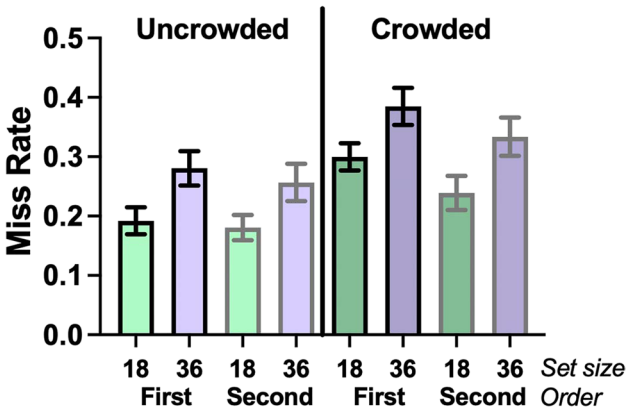

Fig. 3 Error rate for target-present trials (Miss errors). Error bars are ±1 *SEM*

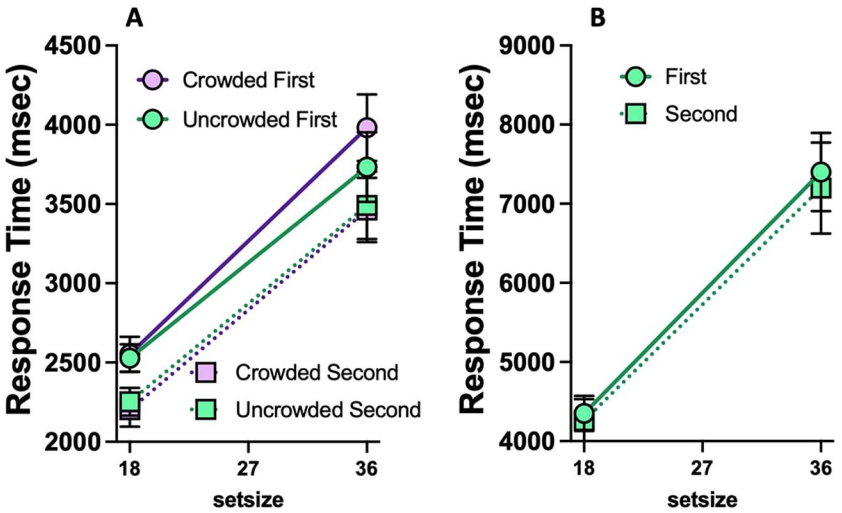

Fig. 2 Response time for target-present trials (**A**) and target-absent trials (**B**). Error bars are ±1 *SEM*

Error rates were quite high. This is probably a function of the effect of online testing (generally poorer performance) and the absence of trial-by-trial feedback. In this experiment, high error rates are actually desirable since they are the primary focus of our statistical analysis. Figure 3 shows the true error rates, but statistical analysis was performed on arcsin-transformed error rates; a method for making error data more normally distributed for statistical analysis (Hogg & Craig, 1995). Again, there was an obvious set size effect, $F(1, 66) = 9.8$, $p < .01$, partial eta-square = 0.09, with somewhat lower error rates on second appearance, $F(1, 66) = 12.8$, $p = .0007$, partial eta-square = 0.16. There was a strong effect of crowding on errors, with an average uncrowded error rate of 22% and an average crowded rate of 31%, $F(1, 66) = 27.6$, $p < .0001$, partial eta-square = 0.29. No interactions reached statistical significance (all ps > .15). Though the crowded and uncrowded targets can be found in the same amount of time, observers are more likely to overlook targets embedded in clusters.

Errors could be looked at in signal detection terms, rather than focusing only on miss errors. However, false-positive (false-alarm) errors are quite rare in this sort of task (except when subjects are performing very poorly). Twenty-five of 34 observers had 0 or 1 false-positive errors on the first repetition of the trials. Twenty-eight had 0 or 1 false-positive errors on the second repetition. There was no significant difference in false-positive errors between first and second repetitions or between crowded and uncrowded stimuli, all paired t tests $t(33) < 1.3$, $p > .2$.

The main purpose of these experiments was to look at the pairs of identical trials and to ask how the chance of missing the target on the second appearance of a display is related to the chance of missing that same target on the first appearance. The data for this analysis are shown in Fig. 4.

The figure plots the probability of missing the target on both its first and second appearances (P12 = P(joint error)) as a function of the probability of missing the target on its first appearance (P1). The deterministic prediction assumes that if the observer misses the first appearance of the target, they will miss the second appearance as well. A simple version of this prediction would generate line of slope = 1. However, observers somewhat improve their performance between first and second appearance, so the deterministic prediction is adjusted to take this effect into account. The deterministic prediction for each observer is min(P1,P2). The simplest stochastic prediction is that P(12), the probability of a joint error, is P(1) multiplied by P(2), shown in red. Each purple outlined square represents the data from one of the 34 observers.

It is clear from Fig. 4 that errors made in both uncrowded and crowded conditions differed from either the stochastic or deterministic predictions (all one-sample t tests, $p < .0001$). To determine the relative contributions of stochastic or deterministic processes, we simulated the experiment. Each simulated observer missed a proportion (P) of targets on the first appearance of those targets. To simulate observers with various levels of skill or diligence, P was randomly chosen between 0 and 0.5. Some fraction (k) of errors were declared to be deterministic, so if a subject missed P targets, P × k of those were deterministic errors. The remainder P × (1 − k) would be stochastic. For the second appearance, if the item had been missed in a deterministic manner on first appearance, it was deemed to be missed on the second appearance. For all remaining items, the chance of missing on the second appearance was P × (1 − k). From these simulated data, we can calculate the probability of missing both items. Figure 5 shows the results for three levels of deterministic errors, plotted in the manner of Fig. 4.

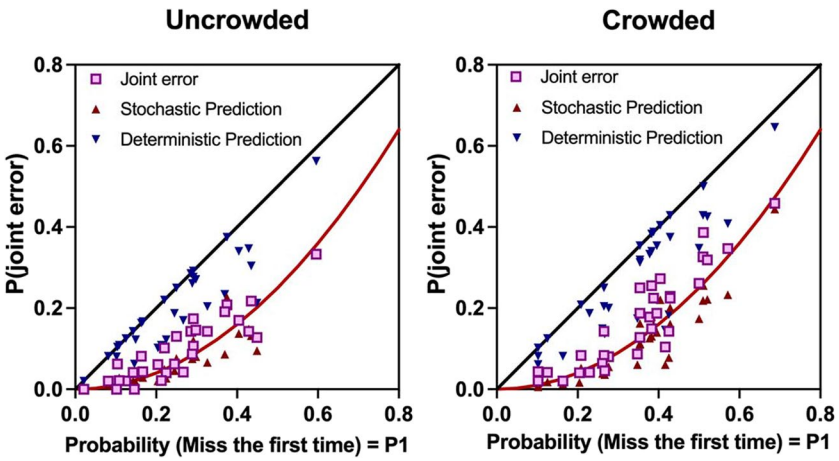

Fig. 4 Probability of missing the target on both its first and second appearances (P12) as a function of the probability of missing the target in its first appearance (P1). Each data point represents one observer. The red triangles show stochastic predictions for each participant, and the blue triangles show the deterministic predictions. Black line shows the simple deterministic prediction: (P12) = (P1). Red line shows the simple stochastic prediction: (P12) = (P1) × (P1). (Color figure online)

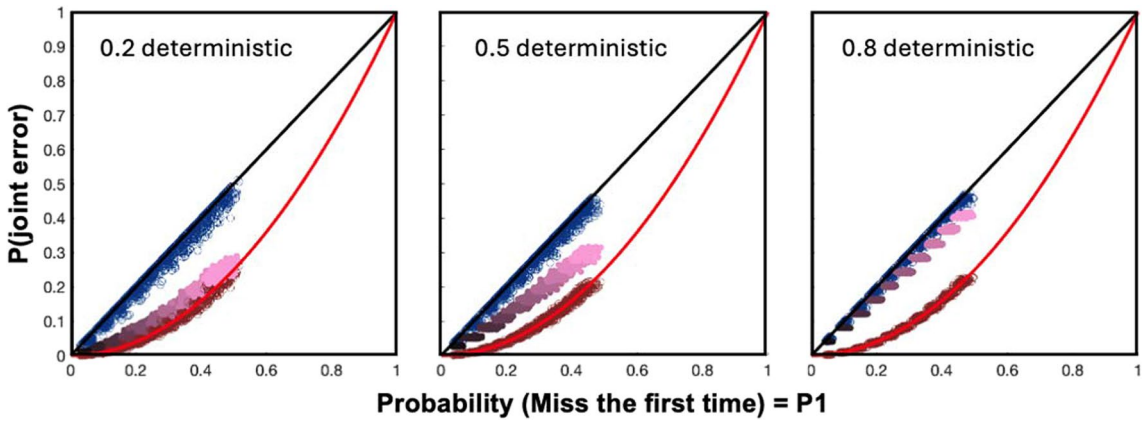

Fig. 5 Simulated plots of the chance of missing a target on both appearances, P(joint error), against the error rate on first appearance, P1. The curved red line and the dark red dots show the pure stochastic prediction. The black line and blue dots show the deterministic predictions. The purple dots show the simulated results for deterministic error proportions of 0.2, 0.5, and 0.8. (Color figure online)

As can be seen, the simulated results fall between the stochastic and deterministic predictions. The average distance between the results for each simulated observer and the stochastic and deterministic predictions for that observer are an almost linear function of the proportion of simulated deterministic errors on first appearance. Thus, we can use the proportion,

$$(\text{P(miss both)} - \text{Stochastic})/(\text{Deterministic} - \text{Stochastic})$$

as an estimate of the relative frequency of the two types of error for each real observer. Figure 6 shows these proportions for each observer in crowded and uncrowded conditions.

The proportions are clearly biased toward the stochastic prediction for most individual observers. The means of 0.20 (uncrowded) and 0.30 (crowded) are significantly less than 0.5 (both one-sample t tests, $p < .0001$). There appear to be more stochastic errors in the uncrowded condition, but this falls short of statistical reliability, $t(31) = 1.928$, $p = .063$. Note there are only 31 observers in this analysis because (P12-S)/(D-S) was undefined for three observers in either or both of the crowded or uncrowded conditions. An informal item analysis did not indicate that specific displays were missed by most observers. Instead, the data indicate that most errors were stochastic. In principle, there could be two types of deterministic errors: individual, where an observer misses a display most of the time and collective, where all observers miss a specific display most of the time. Our design is probably underpowered to assess this distinction with any precision.

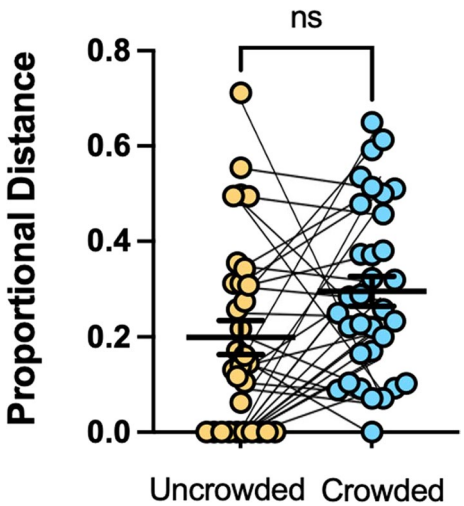

Fig. 6 Proportion of the deterministic errors, based on (P(joint error) − S)/(D − S), where D and S are the deterministic and stochastic predictions for each observer. Values closer to zero reflect a greater proportion of stochastic errors

Discussion

The data indicate that observers make more errors when the target in this *T* among *L* search is located in a crowded cluster of items rather than being more isolated in the display. In and of itself, this is not very interesting. Crowded targets in cluttered displays would be expected to be harder to find. The question for this paper is *why* those additional items are missed. If we had hidden some items behind occluders, they would be missed in an obviously deterministic manner. Are targets, hidden by crowding, missed in a similarly deterministic manner? In fact, this does not appear to be the case. Most of these errors appear to be stochastic, the product of some random process as opposed to being determined by the layout of the display. This could appear somewhat contradictory. Crowding *determines*—at least, to some extent—the error rate in this task, but those errors are largely random in nature. Why would observers randomly miss more targets

in a crowded cluster? This may be a function of incomplete processing of the display. When the eyes land at a location in the visual display, items will be processed within a "functional visual field" (FVF) around the point of fixation. We make 3–4 voluntary saccades and fixations per second. We can select items like these *T*s and *L*s at a rate of one every 50 msec or so. Thus, if more than 5–6 items fall within the FVF, the eyes may move on without all items within the FVF having been selected (Wolfe et al., 2021; Wu & Wolfe, 2022). An unselected item might be successfully selected on another fixation, or it might be missed again. In any case, an item is more likely to be randomly omitted from a crowded region when more items fall in the FVF. Note that this account does not require a commitment to serial selection of items within the FVF. It could be that all items in the FVF are processed but some are not processed adequately. If we assume that inadequate processing is more likely in a crowded FVF, the consequences would be similar to the incomplete serial sampling account. In either case, which item is missed during a fixation might be entirely random or it could be that some configurations lead to more frequent misses. The present results suggest that, if this incomplete processing is the mechanism for these errors, most of items that failed to get selected, failed by chance, producing mostly stochastic errors.

Experiment 2: More difficult *T*s and *L*s

Results of Experiment 1 indicated that most of the false-negative errors were stochastic. Perhaps, if the task were more difficult, observers would develop strategies or bad habits that would make the errors more deterministic (e.g., by always reading from upper left to lower right and quitting before finding targets in that lower right corner. This is a fictional example, not a hypothesis). To assess this, we tested observers with a more difficult version of the T among L search that did not rely on a crowding manipulation. An example of the search task is shown in Fig. 7.

Stimuli and apparatus

Compared with the stimuli in Fig. 1, we further increased the overlap between the line segments making up the *T*s and *L*s. This makes the target more difficult to recognize among distractors, as can be appreciated if you search for the *T* in Fig. 7. Note again, however, that the *T* is unambiguously recognizable, once found. The vertical and horizontal lines of upright targets and distractors were 0.04 screen height. The short segments beyond the intersection of the lines were 0.01 screen height. The orientations of items were randomly chosen from rotations in 30 deg steps from 0 to 330 deg from vertical. All items were placed in a square region of 0.7 × 0.7 screen height, centered on the middle of the screen. The minimum distance between any two letters was constrained to be greater than 0.1 screen height. Set sizes of 18 and 36 were tested. As shown in the figure, the items were white on a mid-gray background. Again, online testing makes precise photometric details unobtainable.

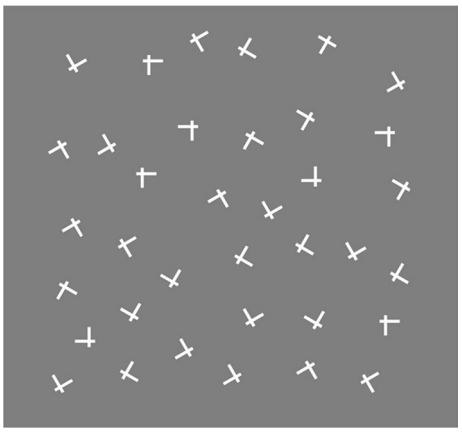

Fig. 7 Hard search for *T* among *L*s. Set size = 36

Participants

In total, 31 participants (14 males, 17 females, $M = 28.19$, $SD = 7.57$, min = 20, max = 54) were recruited online through Prolific, an online crowdsourcing platform. All participants reported normal or corrected to normal vision and accepted an informed consent form presented on the screen before they began the experiment. Participants received 12.9 U.S. dollars in compensation for their participation in the experiment, with payment adjustments made to align with Prolific guidelines. The study took an average of 89 minutes to complete (max = 111, min = 58).

Design and procedure

There were 400 total trials divided between two set size conditions (18 and 36) fully crossed with target status (target absent or present). Two copies of each of 200 unique trials made up the full set of 400 trials. Thus, each combination of set size (18, 36), target presence (yes, no), and repetition (first, second) consisted of 50 trials. These were presented in a random order. Participants went through a four-trial practice session, illustrating the trial types, before they started the experiment. Again, we did not prefilter observers because higher error rates were the goal of the experiment (though we disallowed random performance; see below). There was no trial-by-trial feedback, but participants were shown their accuracy rate after every block of 100 trials. Stimuli were visible until the observer made a response. After the response, observers were allowed a one second

window within which to press another key if they knew that they had made an error (e.g., a simple motor error or a trial where they saw the target while in the process of making an initial response). The next trial began at the end of this one second period.

Data exclusion

Of the 31 participants, 10 were excluded for poor performance, defined as $d' < 1$. This was a hard task and, given our interest in errors, we could choose to accept all participants. However, especially with online studies, there are participants who just do not appear to be doing the task. These are the participants we attempted to remove. For the remaining 21 observers (11 males, 10 females), we removed trials where participants reversed their response, indicating that they had made a motor error. These constituted 1% of trials. Trials with RTs <200 msec were removed as errors of anticipation. RTs greater than 2.5 standard deviations above the mean in each condition were also removed. For any removed trial, we also removed its "twin" from analysis. In total, 4.4% of trials were removed, reflecting the great difficulty of the task.

Results

Figure 8 shows the RT data for correct target-present and target-absent trials for Experiment 2. As would be expected, there was a strong set size effect on the present trials (two-way ANOVA), $F(1, 20) = 94$, $p < .0001$, partial eta-square = 0.82. The slopes of the RT × Set Size functions were 63 msec/item for the first appearance and 74.5 msec/item for the second. This marks the task as relatively difficult and inefficient but indicates that it was not necessary to fixate each item individually in order to identify it. If fixation is required on each stimulus, the target-present slope would be in the range of 125–150 msec/item. There was also a main effect of appearance order with the second appearance of a display producing a substantial decrease in RT (612 msec), $F(1, 20) = 14.2$, $p = .0012$, partial eta-square = 0.41. The interaction of set size and appearance order was not significant ($p = .15$). The same was the case for the absent trials. There was a main effect of set size, $F(1, 20) = 138.5$, $p < .0001$, partial eta-square = 0.87. The main effect of appearance order was marginal, $F(1, 20) = 4.31$, $p < .0509$, partial eta-square = 0.18. The interaction was not significant ($p = .46$).

Figure 9 shows the false negative (Miss) errors. More errors were made with larger set size, $F(1, 20) = 26.4$, $p < .0001$, partial eta-square = 0.57. As in Experiment 1, the ANOVA was conducted with arcsin-transformed data, as is recommend for error rate data. The effect of appearance order was not significant, $F(1, 20) = 1.8$, $p < .20$, partial eta-square = 0.08, but the interaction with set size was $F(1, 20) = 5.5$, $p < .03$, partial eta-square = 0.21. Errors declined on second appearance for set size 18 but not for set size 36. The overall error rate was 43.7%, again, indicating that this was a difficult task. Note that the false-positive (false-alarm) rate is very low with 14 out of the observers having 0 or 1 total false-positive errors. The average false-positive rate was just 1.8% yielding a d' of 2.3 for the task.

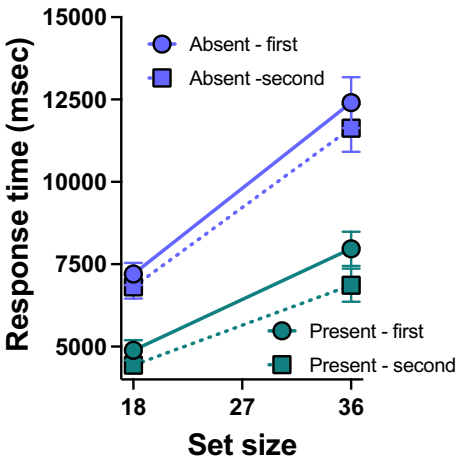

Fig. 8 Response time as a function of set size and first vs second appearance for target-present and target-absent trials. Error bars show ±1 *SEM*

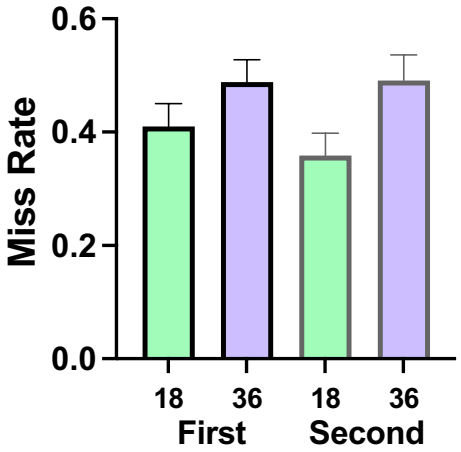

Fig. 9 False negative (Miss) errors as a function of set size and first versus second appearance. Error bars show ±1 *SEM*

As in Experiment 1, the main interest here was in the nature of the errors. Are they stochastic or deterministic? Figure 10 shows the probability of missing both instances of a display plotted against the probability of an error on the first appearance of a display (purple squares). Also shown are the deterministic (min(P1,P2)) (blue) and stochastic (P1 × P2) (red) predictions for each observer.

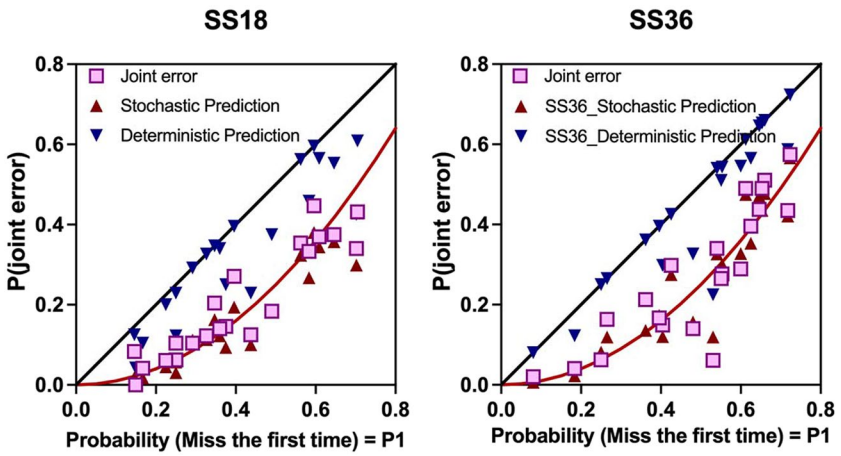

Fig. 10 P(joint error) plotted against the error rate on first appearance. Each purple outlined square represents one observer. Blue triangles show the pure deterministic predictions (min(P1,P2)) for each participant. Red triangles show the stochastic prediction (P1 × P2). Black line shows the simple deterministic prediction: (P12) = (P1). Red line shows the simple stochastic prediction: (P12) = (P1) × (P1). (Color figure online)

On visual inspection, it seems clear that the data lie closer to the stochastic prediction. Statistically, the data differ strongly from the deterministic prediction for both set sizes, $t(20) > 5.63$, $p < .0001$. For the stochastic prediction, the data differs at set size 18, $t(20) = 14.06$, $p < .0001$, but not at set size 36, $t(20) = 1.14$, $p = .26$. Figure 11 shows the relative distance of each observer's data from the stochastic prediction (0) to the deterministic prediction (1), based on the proportion (P12-Stochastic)/(Deterministic-Stochastic). The set size 36 data are closer to the stochastic prediction than the set size 18 data (Fig. 11). This is not technically significant at the $p < .05$ level, $t(20) = 2.085$, $p = .0501$, but it is suggestive. By this estimate, ~80% of errors are stochastic at set size 18 and ~90% at set size 36.

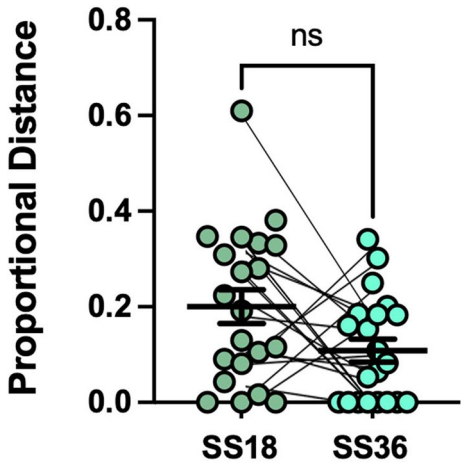

Fig. 11 Relative distance of each observer's data from the stochastic prediction (0) to the deterministic prediction (1)

General discussion

The two experiments reported here add to our understanding of look but fail to see (LBFTS) errors (Hills, 1980; Wolfe, Kosovicheva et al., 2022a). LBFTS errors are those where observers fail to see—or, at least, to report—the presence of stimuli that are clearly visible and recognizable. In the present experiments, these targets were *T*s among *L*s. All the stimuli were presented at high contrast on uniform backgrounds. The *T*s were made somewhat harder to discriminate from the *L*s by having their composite lines overlap to form a + junction, but the target, *T*, was not hard to discriminate from an *L*. Nevertheless, observers missed a sizable number of targets. They made very few false-positive errors. These were errors of omission, not commission. The stimuli were visible until the observer responded, so these are not errors that were caused by lack of time or by information degrading through masking the display. These are quite straightforward LBFTS errors. The errors are akin to typos but, in a sense, more dramatic, since, in the *T* among *L* case, the observers know exactly what they were looking for while a typo can be any member of a broad category.

Our strategy for investigating the cause of these errors is to present a set of search displays twice. With this method, we can distinguish between stochastic and deterministic patterns of error. The critical measure is the proportion of target-present displays that produce errors on both their first and second appearance of the display. If the errors are completely deterministic, then if observers miss the target on the first appearance they should miss it the next time too. If the errors are completely stochastic, then the result on the second appearance would not depend on the

fate of the first appearance. Rather, the chance of missing the target both times would be the product of the chance of missing it the first time and the chance of missing it the second time.

The results of both experiments show a strong bias toward stochastic errors, though the data do not fit the stochastic prediction perfectly. In one case (Exp. 2, set size 36), the data and the prediction of the stochastic model were not statistically different. In all other cases, there were more joint errors than a pure stochastic model would predict. The distance between the stochastic and deterministic predictions constitutes an approximately linear scale that can be used to partition errors into stochastic and deterministic fractions. The highest proportion of deterministic errors was found in the crowded condition of Experiment 1. Even there, only 30% of errors appeared to be deterministic. The uncrowded condition produced 20% errors—not significantly different. In Experiment 2, 20% of set size 18 errors were deterministic against 11% of the set size 36 errors—again, not significantly different. In Experiment 1, there were significantly more errors, overall, in the crowded condition. We hypothesize that the higher overall error rate in the crowded condition arises from a failure to process all the items within a functional visual field (FVF) around the current point of fixation. In a clump, there will be more items on average, increasing the chance that an individual target item will not be adequately attended. Moreover, targets in clumps are more likely to be subject to crowding effects (Whitney & Levi, 2011). Potentially, crowding effects might account for some of the deterministic errors. However, crowding will interact with the location of fixation. Relatively random fixations might allow crowding to contribute to stochastic errors rather than deterministic errors because on one random fixation near a clump, the target might be more peripheral and more likely to be missed. That said, some displays might be more conducive to missing a target. The argument for deterministic displays would be stronger if it were the case that all or nearly all observers missed the target in some specific displays. Analyses of errors as a function of specific display did not reveal any that were particularly prone to errors. The patterns of paired errors look quite idiosyncratic. However, this study was probably underpowered to detect any but the largest effects of this sort, leaving the topic for future research. Moreover, if the density of items in clumps was a cause of more deterministic errors in Experiment 1, we might expect more deterministic errors in the larger set size of Experiment 2; this was not the case.

As a direction for future study, it might be valuable to repeat these experiments while eye tracking participants, something that could not be done with online testing. Eye tracking would allow us to distinguish between what Kundel and colleagues (1978) call "search errors," where the eyes never land in the vicinity of the target, and "recognition" errors, where the eyes fixate on or near the target but then refixate elsewhere within a fraction of a second. Understanding repeated errors would benefit from seeing where observers fixated. It would also be valuable to repeat these experiments with different types of stimuli. In particular, it would be valuable to know if similar results would be obtained with more naturalistic and or socially significant stimuli.

Why is it worth trying to understand the roots of look but fail to see errors? As noted at the outset, some of these LBFTS errors occur in settings where the failure can be dangerous in one way or another. If we understood why we made the errors, we would have a better chance of reducing them in settings where it is important to do so. Consider the case of medical image perception (Samei & Krupinski, 2018). We know that even skilled radiologists make errors (Berlin, 2007; Goddard et al., 2001; Waite et al., 2016), and that many errors are "retrospectively visible" (Boyer et al., 2004; Nodine et al., 2001). Retrospectively visible errors are those where the finding is clearly visible when pointed to, after the fact. It is usually obvious that the radiologist looked at the image. In eye-tracking studies, it can frequently be seen that the eyes fixated on or near the missed item (Kundel et al., 1978; Wolfe et al., 2021), making these quite clearly LBFTS errors.

Obviously, these LBFTS errors can have negative impact on patients. These are also the sorts of "perceptual errors" (Berlin & Hendrix, 1998) that result in malpractice charges (Duszak & Robinson, 2022). If such cases get to trial, they often go poorly for the radiologist because juries and the legal system more generally have trouble understanding how an LBFTS error can be anything but "negligence." If at least some LBFTS errors in radiology are similar in kind to the errors studied here, it could be argued that it is not reasonable to consider them evidence of negligence (Bruno et al., 2024). They might be better considered to be examples of "normal blindness" (Wolfe, Kosovicheva et al., 2022a); a problem that we should try to mitigate, not litigate.

Is there evidence that radiologist errors share any common mechanisms with the largely stochastic errors produced in our *T* among *L* searches? We suggested, above, that one path to stochastically missed *T*s was the failure to process all of the items inside a functional visual field around the current point of fixation. Evidence for this comes from eye-tracking studies. In standard search, observers reliably fixate on the target once they locate it. We have found that, even when the eyes are fixated right next to the target, the target is only fixated immediately thereafter about 50% of the time. The other 50% of saccades move the eyes elsewhere in the field (Wu & Wolfe, 2022). A very similar pattern of results is seen when mammographers search for signs of cancer in a mammogram (Wolfe et al., 2021). When the eyes are right next to the lesion (~1 deg away), there is only a 50% chance that the next fixation lands on the lesion. Of course, radiologists do not miss 50% of cancers. They make many fixations while examining the image,

and, if they do not notice the lesion the first time, they will probably find it when the eyes fixate in the right neighborhood after a later saccade. Still, on the occasions when they do not get back to the right location, they will miss a target that could have been detected. Thus, there is some reason to believe that a common cause could lie behind LBFTS errors in simple search and in medical image perception. Similar accounts could be offered for some driving errors, errors in security screening, and other important and error-prone search tasks.

Does this account of, at least some, LBFTS errors offer any hope for reducing these errors? In fact, the answer is "yes." Recall that the bulk of the errors in these experiments appear to be stochastic in nature. The chance of missing the target in its second appearance was largely, though not completely, independent of the chance of missing the same target the first time. That suggests that one way to reduce these errors is to search twice. Consider, for example, the set size 18 data for Experiment 2. Observers missed 41% of targets on the first appearance, 36% on the second appearance, but only 20% on both appearances. This is somewhat more than the $0.41 \times 0.36 = 15\%$ predicted by complete independence but there is a reduction in errors that could be obtained by looking twice.

There are many ways to look twice. The same observer can look twice, as in the present experiments. More typically in radiology, two experts would look at a case. This is common practice in Europe, where so-called double reading clearly reduces false-negative errors (e.g., Taylor-Phillips et al., 2018). There are variations on double reading. For instance, does the second reader know what the first reader reported or not? That is, are the two readers truly independent? Double reading is not common in the USA. The obvious problem is that two independent reads will take twice as long, will markedly increase expense, and would require more radiologist time when it is already hard to adequately staff radiology clinics. Moreover, since expert radiologists miss very few clearly visible cancers, the benefits of double reading will be small in absolute terms. Still, if your errors are stochastic, an extra set of eyeballs is an obvious intervention.

These days, those eyeballs would not need to be human. Advances in artificial intelligence make it reasonable to propose AI as the second reader (e.g., Dahlblom et al., 2023; Koch et al., 2023). A skilled AI may not replace the human reader, at least not in the short term, but, again, if errors are largely stochastic, two independent readers will catch most of each other's random errors.

In sum, we evaluated the nature of errors in a simple but demanding search task. In the two versions tested here, observers made substantial numbers of false-negative errors. Our analysis indicates that the largest proportion of these errors were stochastic in nature. This adds to our understanding of LBFTS errors and points to the potential of double reading to improve the performance of a system, even if the individual continues to make errors.

Data transparency/Open practices statement Both Experiments 1 and 2 were preregistered on the Open Science Framework, and all study data and code can be accessed at:
Experiment 1: https://osf.io/swyjf/
Experiment 2: https://osf.io/vxfm7/

Funding Jeremy Wolfe and Ava Mitra were supported by National Institute of Health-National Eye Institute: EY017001, National Science Foundation: 2146617, and National Institute of Health-National Cancer Institute: CA207490. Johan Hulleman was supported by UK Research and Innovation (UKRI) grant ES/X000443/1. Wentao (Taylor) Si was supported by Bates College.

Declarations

Ethics approval and consent to participate The research was approved by the Mass General Brigham IRB Protocol #: 2009P001253 and all participants gave consent at the start of the online experiments.

Consent for publication All authors and participants consent to publication of this work.

Competing interests All authors declare no conflict of interest.

References

Berbaum, K. S., Franken, E. A., Caldwell, R. T., Shartz, K., & Madsen, M. (2019). Satisfaction of search in radiology. In E. Samei & E. A. Krupinski (Eds.), *The handbook of medical image perception and techniques* (2nd ed., pp. 121–166). Cambridge University Press.

Berlin, L. (2007). Radiologic errors and malpractice: A blurry distinction. *American Journal of Roentgenology, 189*(3), 517–522. https://doi.org/10.2214/ajr.07.2209

Berlin, L., & Hendrix, R. W. (1998). Perceptual errors and negligence. *American Journal of Roentgenology, 170*(4), 863–867. https://doi.org/10.2214/ajr.170.4.9530024

Boyer, B., Hauret, L., Bellaiche, R., Graf, C., Bourcier, B., & Fichet, G. (2004). Retrospectively detectable carcinomas: Review of the literature [Review]. *Journal De Radiologie, 85*(12), 2071–2078. https://doi.org/10.1016/s0221-0363(04)97784-0

Bruno, M. A., Krupinski, E. A., Bunce, S. C., Baird, G., Mills, C., Karunanayaka, P. R., Howard, E., Chang, R., Cottrill, R., Jump, S., Krishnankutty, S., & Mosher, T. J. (2024). Neurocognitive Mechanism of Radiologists' Perceptual Errors: Results of Preliminary Studies. *bioRxiv*. https://doi.org/10.1101/2024.04.29.591746

Cain, M. S., Adamo, S. H., & Mitroff, S. R. (2013). A taxonomy of errors in multiple-target visual search. *Visual Cognition, 21*(7), 899–921. https://doi.org/10.1080/13506285.2013.843627

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology, 30*, 39–78.

Chun, M., & Jiang, Y. (1998). Contextual cuing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36*, 28–71.

Dahlblom, V., Dustler, M., Tingberg, A., & Zackrisson, S. (2023). Breast cancer screening with digital breast tomosynthesis: comparison of different reading strategies implementing artificial intelligence. *European Radiology, 33*(5), 3754–3765. https://doi.org/10.1007/s00330-022-09316-y

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. Academic Press.

Duszak, R., Jr., & Robinson, J. (2022). Malpractice litigation: The elephant in the reading room. *Journal of the American College of Radiology.* https://doi.org/10.1016/j.jacr.2022.05.001

Goddard, P., Leslie, A., Jones, A., Wakeley, C., & Kabala, J. (2001). Error in radiology. *British Journal of Radiology, 74*(886), 949–951. https://doi.org/10.1259/bjr.74.886.740949

Greenlee, E. T., DeLucia, P. R., & Newton, D. C. (2022). Driver vigilance decrement is more severe during automated driving than manual driving. *Human Factors, 66*(2), 574–588. https://doi.org/10.1177/00187208221103922

Hanna, T. N., Zygmont, M. E., Peterson, R., Theriot, D., Shekhani, H., Johnson, J.-O., & Krupinski, E. A. (2018). The effects of fatigue from overnight shifts on radiology search patterns and diagnostic performance. *Journal of the American College of Radiology, 15*(12), 1709–1716. https://doi.org/10.1016/j.jacr.2017.12.019

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8*, 150–150. https://doi.org/10.3389/fnins.2014.00150

Hills, B. L. (1980). Vision, visibility, and perception in driving. *Perception (London), 9*(2), 183–216. https://doi.org/10.1068/p090183

Hogg, R. V., & Craig, A. T. (1995). *Introduction to mathematical statistics* (5th ed.). Prentice-Hall.

Koch, H. W., Larsen, M., Bartsch, H., Kurz, K. D., & Hofvind, S. (2023). Artificial intelligence in BreastScreen Norway: A retrospective analysis of a cancer-enriched sample including 1254 breast cancer cases. *European Radiology, 33*(5), 3735–3743. https://doi.org/10.1007/s00330-023-09461-y

Kouskos, E., Markopoulos, C., Mantas, D., Revenas, K., Antonopoulou, Z., Kontzoglou, K., & Gogas, J. (2004). Missed cancers on mammograms: Causes and measures of prevention. *European Journal of Gynaecological Oncology, 25*(2), 230–232.

Krupinski, E. (2010). Reader fatigue interpreting mammograms. In E. A. J. Marti (Ed.), *International workshop on digital mammography* (vol. LNCS 6136, pp. 312–318). Springer.

Kundel, H., L. (2007). How to minimize perceptual error and maximize expertise in medical imaging. *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment, 6515*. https://doi.org/10.1117/12.718061

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology, 13*(3), 175–181. https://journals.lww.com/investigativeradiology/abstract/1978/05000/visual_scanning,_pattern_recognition_and.1.aspx. Accessed 2 Aug 2004.

Li, A., Wolfe, J. M., & Hulleman, J. (2024). Errors in visual search: Are they stochastic or deterministic? *Cognitive Research: Principles and Implications, 9*, 15. https://doi.org/10.1186/s41235-024-00543-z

Meuter, R. F. I., & Lacherez, P. F. (2016). When and why threats go undetected: Impacts of event rate and shift length on threat detection accuracy during airport baggage screening. *Human Factors, 58*(2), 218–228. https://doi.org/10.1177/0018720815616306

Nartker, M., Firestone, C., Egeth, H., & Phillips, I. (2023). Six ways of failing to see (and why the differences matter). *i-Perception, 14*(5), 20416695231198760. https://doi.org/10.1177/20416695231198762

Nodine, C. F., Mello-Thoms, C., Weinstein, S. P., Kundel, H. L., Conant, E. F., Heller-Savoy, R. E., Rowlings, S. E., & Birnbaum, J. A. (2001). Blinded review of retrospectively visible unreported breast cancers: An eye-position analysis. *Radiology, 221*(1), 122–129. https://doi.org/10.1148/radiol.2211001507

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision [Hypothesis & Theory]. *Frontiers in Psychology, 3*, 13. https://doi.org/10.3389/fpsyg.2012.00013

Rubinstein, J. (2020). Divergent response-time patterns in vigilance decrement tasks. *Journal of Experimental Psychology: Human Perception and Performance., 46*(10), 1058–1076. https://doi.org/10.1037/xhp0000813

Ruckmick, C. A. (1926). On overlooking familiar objects. *The American Journal of Psychology, 37*(4), 631–632. https://doi.org/10.2307/1414949

Samei, E., & Krupinski, E. A. (2018). *The handbook of medical image perception and techniques* (2nd ed.). Cambridge University Press.

Schall, J. D. (2019). Accumulators, neurons, and response time. *Trends in Neurosciences, 42*(12), 848–860. https://doi.org/10.1016/j.tins.2019.10.001

Sharpe, B. T., Smith, M. S., Williams, S. C. R., Talbot, J., Runswick, O. R., & Smith, J. (2023). An expert-novice comparison of lifeguard specific vigilance performance. *Journal of Safety Research, 87*, 416–430. https://doi.org/10.1016/j.jsr.2023.08.014

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059–1074.

Taylor-Phillips, S., Jenkinson, D., Stinton, C., Wallis, M. G., Dunn, J., & Clarke, A. (2018). Double reading in breast cancer screening: Cohort evaluation in the CO-OPS trial. *Radiology, 287*(3), 749–757. https://doi.org/10.1148/radiol.2018171010

Thomson, D. R., Besner, D., & Smilek, D. (2015). A resource-control account of sustained attention: Evidence from mind-wandering and vigilance paradigms. *Perspectives on Psychological Science, 10*(1), 82–96. https://doi.org/10.1177/1745691614556681

Titchener, E. B. (1924). The overlooking of familiar objects. *The American Journal of Psychology, 35*(2), 304–305. https://doi.org/10.2307/1413844

Waite, S., Scott, J., Gale, B., Fuchs, T., Kolla, S., & Reede, D. (2016). Interpretive error in radiology. *American Journal of Roentgenology, 208*(4), 739–749. https://doi.org/10.2214/AJR.16.16963

Whitney, D., & Levi, D. M. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences, 15*(4), 160–168. https://doi.org/10.1016/j.tics.2011.02.005

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychological Bulletin & Review, 28*, 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Wolfe, J. M., Kosovicheva, A., & Wolfe, B. A. (2022a). Normal blindness: When we look but fail to see. *Trends in Cognitive Science, 26*, 809–819. https://doi.org/10.1016/j.tics.2022.06.006

Wolfe, J. M., Lyu, W., Dong, J., & Wu, C.-C. (2022b). What eye tracking can tell us about how radiologists use automated breast ultrasound. *Journal of Medical Imaging (Bellingham), 9*(4), 045502. https://doi.org/10.1117/1.JMI.9.4.045502

Wolfe, J. M., Wu, C. C., Li, J., & Suresh, S. B. (2021). What do experts look at and what do experts find when reading mammograms? *Journal of Medical Imaging (Bellingham), 8*(4), 045501. https://doi.org/10.1117/1.Jmi.8.4.045501

Wu, C. C., & Wolfe, J. M. (2022). The functional visual field(s) in simple visual search. *Vision Research, 190*, 107965. https://doi.org/10.1016/j.visres.2021.107965

Zenger, B., & Fahle, M. (1997). Missed targets are more frequent than false alarms: A model for error rates in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 23*(6), 1783–1791.

Zhang, X., Huang, J., Yigit-Elliott, S., & Rosenholtz, R. (2015). Cube search, revisited. *Journal of Vision, 15*(3), 9–9. https://doi.org/10.1167/15.3.9