



Taking prevalence effects on the road: Rare hazards are often missed

Anna Kosovicheva¹ · Jeremy M. Wolfe^{2,3} · Benjamin Wolfe¹

Accepted: 25 July 2022
© The Psychonomic Society, Inc. 2022

Abstract

Previous work has shown that, in many visual search and detection tasks, observers frequently miss rare but important targets, like weapons in bags or abnormalities in radiological images. These prior studies of the low-prevalence effect (LPE) use static stimuli and typically permitted observers to search at will. In contrast, many real-world tasks, like looking for dangerous behavior on the road, only afford observers a brief glimpse of a complex, changing scene before they must make a decision. Can the LPE be a factor in dynamic, time-limited moments of real driving? To test this, we developed a novel hazard-detection task that preserves much of the perceptual richness and complexity of hazard detection in the real world, while allowing for experimental control over event prevalence. Observers viewed brief video clips of road scenes recorded from dashboard cameras and reported whether they saw a hazardous event. In separate sessions, the prevalence of these events was either high (50% of videos) or low (4%). Under low prevalence, observers missed hazards at more than twice the rate observed in the high-prevalence condition. Follow-up experiments demonstrate that this elevation of miss rate at low prevalence persists when participants were allowed to correct their responses, increases as hazards become increasingly rare (down to 1% prevalence) and is resistant to simple cognitive intervention (participant prebriefing). Together, our results demonstrate that the LPE generalizes to complex perceptual decisions in dynamic natural scenes, such as driving, where observers must monitor and respond to rare hazards.

Keywords Visual attention · Driving · Prevalence · Road hazards

Our survival hinges on our ability to detect and respond to dangerous events. Most of the time, these critical situations are rare. Driving is an important example, where hazards requiring a rapid response are relatively infrequent (e.g., the vehicle ahead of you suddenly braking; Dingus et al., 2016). The prevalence of these events can vary considerably across environments and can take many forms. For example, the sudden appearance of pedestrians will be more common on a busy urban street than on a lonely highway. Here, we examine how the relative prevalence of road hazards affects our ability to respond to them. This can be seen as one way to examine prevalence effects in dynamic scenes more generally.

Our inability to notice infrequent events has been a major topic of research in cognitive psychology. Work in visual search has demonstrated the low-prevalence effect, or LPE, (Horowitz, 2017; J. M. Wolfe et al., 2005) in which observers frequently miss rare but important targets, including threats in luggage (J. M. Wolfe et al., 2013; J. M. Wolfe et al., 2007) and abnormalities in medical images (Evans et al., 2011; Evans et al., 2013). In these static visual search tasks, observers become less willing to say that targets are present as targets become rarer. Within a signal detection framework, this has been shown to reflect a shift toward a more conservative response criterion (c) rather than a change in sensitivity, or d' (J. M. Wolfe et al., 2007). Similar results are also well established in other domains. For example, decision theory defines an optimal criterion based on the relative frequency of signal and noise events, combined with the payoffs associated with different responses (D. M. Green & Swets, 1966). In the vigilance literature, when participants must respond to events over an extended period (N. H. Mackworth, 1948; J. Mackworth, 1970), they show worse performance as

✉ Anna Kosovicheva
a.kosovicheva@utoronto.ca

¹ Department of Psychology, University of Toronto
Mississauga, 3359 Mississauga Road, Mississauga,
ON L5L 1C6, Canada

² Brigham and Women's Hospital, Boston, MA, USA

³ Harvard Medical School, Boston, MA, USA

signal frequency decreases (Baddeley & Colquhoun, 1969; Broadbent & Gregory, 1965; Colquhoun, 1961).

There are situations where the reverse phenomenon is seen; where observers adopt a more *liberal*, rather than a more conservative criterion, as targets become rare. In particular, Levari et al. (2018) reported the phenomenon of prevalence-induced concept change (PICC), in which participants' definition of what constitutes a target expands as targets become increasingly rare. For example, when categorizing colored patches as blue or purple, patches that were ambiguous were *more* likely to be categorized as blue when blue targets were rare. More recently, these findings have been shown to depend on the feedback that participants receive on each trial (Lyu et al., 2021). Without any feedback on the accuracy of their responses, observers adopt a more liberal criterion (i.e., "I keep saying 'purple,' so perhaps I should say 'blue' more often). However, when provided with trial-by-trial feedback on response accuracy, they learn the true prevalence rate of the target and adopt a more conservative criterion.

How would target prevalence influence detection when the target was a transient event encountered in a dynamic scene, like a road hazard in a driving context? Drivers must perceive, comprehend, and respond to hazards within 1,500–2,000 ms (M. Green, 2000). Our ability to do so is likely to rely on our ability to perceive the gist of a scene (Oliva, 2005) and to classify the kind of scene we are seeing based on a single brief glimpse (Greene & Oliva, 2009). On the road, we must do so while we move through the scene and it moves around us, likely making this perceptual task harder. Yet observers are able to accurately detect road hazards having only glimpsed a dynamic scene for 200–300 ms (B. Wolfe et al., 2020), suggesting that the duration of visual input required for initial hazard detection is brief, relative to the time required to plan and execute a response.

Road hazard detection is made more complex by the heterogeneity of hazards in type and size (e.g., pedestrians, animals, vehicles, objects), temporal unpredictability, and variation in visual field location (i.e., hazards are generally on or near the road, but could, for instance, drop from above). This heterogeneity could actually help drivers detect infrequent hazards. Adding visual variability has been shown to reduce vigilance-related performance decrements, during simulated driving in monotonous environments (Thiffault & Bergeron, 2003) and with simple laboratory stimuli (Thomson et al., 2015). In contrast, other work has shown that as vigilance tasks become more cognitively demanding, vigilance-related performance decrements are larger (e.g., Parasuraman & Davies, 1977; Warm et al., 2008).

Previous work suggests that the LPE may be a factor in the detection of on-road targets. Drawing on studies of visual search under low prevalence conditions, Beanland and colleagues asked whether target prevalence could manipulate

search performance in a driving simulator (Beanland et al., 2014). Participants were asked to report every instance of a bus or motorcycle they saw while driving a route in the simulator, under conditions where buses or motorcycles were common or uncommon. Their participants were slower to detect low-prevalence targets, regardless of vehicle type. For example, when buses were prevalent and motorcycles were rare, observers were slower to identify the presence of a motorcycle.

The task of detecting and responding to road hazards as a general class of event is quite different from tasks in which the observer must detect a specific vehicle type (e.g., motorcycles). In fact, one might imagine that the diverse range of hazards might make an LPE more likely. In the experiments presented here, we show that there is a driving hazard LPE. It involves a shift in response criterion, apparently similar to the mechanisms involved in low prevalence visual search (J. M. Wolfe et al., 2007). This LPE is not due to motor errors (Fleck & Mitroff, 2007). And importantly, this effect is resistant to a participant briefing and is not reduced when participants are made aware of the effect.

There are certain challenges (ethical and otherwise) to studying prevalence effects in real-world driving situations, since near collisions are relatively rare. Moreover, experimental design would be additionally complicated in an on-road setting by the fact that hazard prevalence covaries with other environmental factors, like time of day and weather, to say nothing of safety considerations. A simulator would not provide the diversity of real-world hazards. Viewing driving videos, while there is some risk that the observer may adopt a passive viewing stance, may be the best of these options, balancing realism with practical constraints. Here we make use of a set of real road videos of naturally occurring hazards (B. Wolfe et al., 2020). This approach allows us to include much of the richness, visual complexity, and perceptual variability from the road, while still allowing us to maintain safety and experimental control over event prevalence.

Methods

In each of five experiments, we asked participants to detect hazardous situations (e.g., sudden braking, pedestrians, animals on the road; see Fig. 1) in brief video clips (333 ms) of real road scenes. First, Experiment 1 compares performance at hazard rates of 50% versus 4%. Results show an elevated proportion of missed targets at low prevalence, replicating a classic LPE (J. M. Wolfe et al., 2005) in a hazard-detection paradigm. In Experiments 2 and 3, we show that this effect is dependent on receiving trial-wise feedback (similar to Lyu et al., 2021). In Experiment 4, where participants could correct their initial responses (Fleck & Mitroff, 2007), we eliminate motor errors as an explanation for our effect. Finally, in

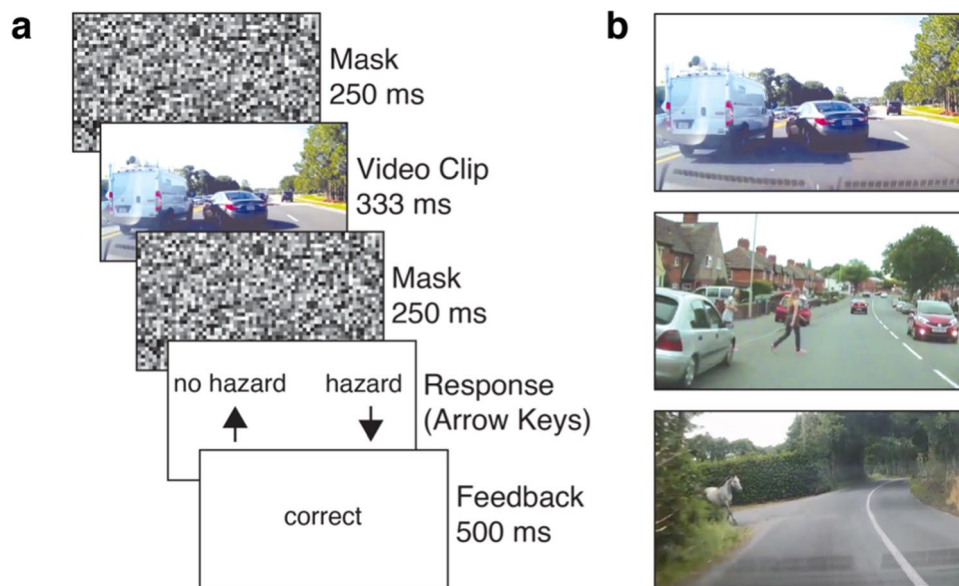


Fig. 1 **a** Sequence of events in each trial. Participants were shown a random noise mask for 250 ms, followed by a 333-ms video clip of a road video recorded from a dashboard camera. Following a post-stimulus mask (250 ms), participants reported whether the video contained a hazard by pressing the up or down arrow key. Participants received feedback regarding the accuracy of their response (all Experiments except 2 and 3; see Methods). The frequency of hazard-

present videos was varied between conditions (rates of 50% vs. 4% in Experiments 1–4, and 1% vs. 10% in Experiment 5). **b** Examples of hazardous road events used in the experiments (e.g., unsafe driving by other vehicles in the lane of travel and pedestrians or animals in the roadway). All videos showed real road scenes and covered a diverse range of settings (urban, highway, rural), times of day, and weather conditions

Experiment 5, we take a step toward a more realistic hazard prevalence rate (10% vs. 1%), demonstrating that the effect scales as prevalence declines. In this experiment, we also show that these effects persist when observers are informed about the prevalence effect, suggesting a resistance to pre-briefing and driver instruction.

All experiments shared a common experimental procedure (described in Procedure, below), but varied in their particulars. Changes to this common procedure, like manipulations to the feedback provided to participants or changes to the hazard prevalence in a given experiment, are described in the Additional Experiments section.

Participants

Participants were recruited online through Prolific (www.prolific.co), a platform for human participant research. There were 16 participants per experiment in Experiments 1–4 and 16 in each of two conditions of Experiment 5 for a total of 32. All participants were, by self-report, between the ages of 20 and 35, in possession of a valid driving license, had corrected to normal visual acuity, and were residents in Canada, the United States of America, or the United Kingdom. To ensure that a new set of participants was tested in each experiment, Prolific users who had previously completed one experiment were excluded from future experiments. The mean age across all experiments was 27.0

years ($SD = 4.4$ years). The sample size of 16 was determined from the effect size measured in a pilot study with 11 observers, which indicated that a minimum of seven participants were necessary to detect a significant difference in criterion between the low (4%) and high (50%) prevalence conditions at 95% power ($d_z = 1.71$, $\alpha = .05$).

Procedures were approved by the Research Ethics Board at the University of Toronto, and all participants provided informed consent prior to participating in the experiment. Each experiment consisted of two sessions, each lasting approximately 30 minutes, and participants were paid £4.50 per session on Prolific, with a completion bonus of £8.50 for finishing both sessions.

Each experiment was pre-registered on the Open Science Framework (OSF) online prior to data collection. The preregistration information, and all materials for this study (stimuli, experimental code, anonymized data, and analysis scripts) are available from OSF (<https://osf.io/r9uk7>).

Stimuli

The experiment was conducted online and programmed using PsychoPy/PsychoJS (Version 2020.2; Peirce et al., 2019), and hosted online on Pavlovia, based on results showing that PsychoPy/PsychoJS was the lowest latency platform for online studies (Bridges et al., 2020). Participants were required to complete the study on a desktop

or laptop computer (i.e., the experiment was disabled on mobile and tablet devices).

The videos used in the study were created from an extended version of the Road Hazard Stimulus set (available at <https://osf.io/uq6pc/>). This stimulus set consists of forward-facing road videos recorded from dashboard cameras, sourced from YouTube, in collaboration with the Moments project at MIT (Monfort et al., 2019; see B. Wolfe et al., 2020, for full details of the stimulus set). Briefly, the video set consists of a wide variety of road hazards (including uncontrolled objects, pedestrians, animals, and other vehicles) captured in a range of road environments (e.g., highway, city streets, rural roads), weather conditions, and times of day. Hazards were defined as situations that required an immediate driver response, and videos were selected to maximize hazard variability. The videos had been previously annotated for the time of hazard onset as well as the first time of visible driver response (as detailed in B. Wolfe et al., 2020). To control for environmental factors, a corresponding set of matched nonhazard videos were taken from epochs in the hazardous source video at least 10 s prior to hazard onset, whenever possible. To reach the number of videos required for this study, the original Road Hazard Stimulus set was supplemented with additional videos from online sources (e.g., YouTube), which were processed and annotated according to the same procedure and specifications as the original set. This extended version of the Road Hazard Stimulus set consisted of 436 video clips containing hazardous events (8 s in duration) and 316 nonhazard videos (ranging from 1.4 to 22.7 s in duration). This heterogeneous stimulus set allows us to ask general cognitive questions (in this case, about the detection of rare events) using dynamic natural scenes. If one wanted to draw conclusions about specific road environments (e.g., specific types of infrastructure or other factors like driving on the left vs. the right), a more focused stimulus set would be required.

Finally, to produce the videos used in the experiment, this stimulus set was broken up into segments that were 333-ms long. For the hazard-present condition, these video segments were taken from the 333 ms immediately preceding the annotated onset of the driver response in each video. For the hazard-absent condition, these video segments were taken from 333-ms epochs every 3 s in each clip. The final set consisted of 436 unique hazard-present and 925 unique hazard-absent clips. The audio was removed from each video, and all videos had a resolution of $1,280 \times 720$ pixels and a frame rate of 30 frames per second. Sample videos used in the experiment can be accessed online (<https://osf.io/r9uk7/files/>).

Procedure

Procedures were adapted from a previous hazard-detection task (B. Wolfe et al., 2020). At the onset of each trial,

observers were shown a random noise mask ($1,280 \times 720$ pixels) for 250 ms, consisting of a grid of 36×64 squares, 20 pixels high each, with a random grayscale intensity from 0 to 255. Following the mask, observers viewed the road video for 333 ms (10 video frames at 30 fps). This duration was selected to be longer than the duration required for 80% hazard-detection accuracy for the majority of observers, based on a separate online replication of the study described in (B. Wolfe et al., 2020) with age-matched observers. In this previous study, the mean duration threshold was 228 ms. Here, we added an additional 105 ms of viewing time to this threshold (for a total duration of 333 ms), presenting stimuli at approximately 1.5 times the mean duration threshold for this task. If one uses much longer durations, the videos begin to include the onset of the driver's response. This response produces cues to the presence of a hazard (e.g., change in optic flow from the driver suddenly braking, or from the vehicle swerving) that would confound our measure of hazard detection with a measure of the detection of the response to the hazard. Extending the clip earlier in time merely adds "normal" stimulus time to the clip. It does help to set a context, but this additional contextual information seems to have a negligible effect on hazard-detection tasks using the same video stimuli (B. Wolfe et al., 2020).

In the high-prevalence condition, 50% of the road videos had a hazard (220 out of 440), as used in previous studies (B. Wolfe et al., 2020), and in the low-prevalence condition, 4% of the road videos had a hazard (20 out of 500). Hazard and nonhazard videos were randomly interleaved, and a given hazard or nonhazard video was selected randomly on each trial from the set of videos which had not been shown to the observer. Following a second (poststimulus) mask, shown for 250 ms, observers reported whether or not a hazard was present using the up and down arrow keys (up: hazard absent, down: hazard present). Hazardous situations were defined to participants as events "that would require your response in an on-road situation." Participants were provided with examples (e.g., sudden braking, near collisions, pedestrians, animals) at the beginning of the experiment, but were informed that hazards would not be limited to these examples. After indicating their response, participants received feedback on their accuracy after each trial, where they were shown a screen with the text "correct" or "incorrect" for 500 ms. The next trial began automatically after a gray screen shown for 500 ms, plus any additional time required for loading the next video. No road video was seen more than once by the same participant.

To familiarize observers with the hazards that would be shown, at the beginning of each session, participants completed 40 practice trials at 50% prevalence and a 1,000-ms video duration with the same trial-wise feedback. Following the practice, for low- and high-prevalence conditions, respectively, participants were informed that "unlike the practice, hazards will now

be relatively rare,” or “hazards will appear frequently, like the practice,” but were not given any further information as to their frequency. Participants were continuously shown a progress bar above the video indicating the proportion of the session completed. Every 60 trials, participants were given the opportunity to take a break and were shown a screen indicating their overall percentage accuracy up until that point.

As a data quality measure, and to remove potential automated (i.e., nonhuman) responders, each session included 40 catch trials, randomly interleaved, in which observers were given an orthogonal scene discrimination task. These trials followed the same procedure as the main task, except the road video was replaced with a 1,000-ms video clip of a nondriving scene, and participants were required to report whether the scene was indoor or outdoor. These clips were sourced from YouTube and public-domain stock footage websites (e.g., [Pexels.com](https://www.pexels.com)), and composed of a range of everyday settings (e.g., living room, beach, forest, library) and activities (e.g., cooking, hiking, gardening). After the video, participants were shown a screen that indicated the instructions for catch trials (i.e., to press the up arrow key to indicate an outdoor video, down arrow key for an indoor video). Each session included 20 indoor and 20 outdoor videos, and each video was randomly selected from a full set of 80 videos across the two sessions.

Each participant completed the low- and high-prevalence conditions (580 and 520 trials, respectively, including the 40 catch trials and 40 practice trials) in two sessions on different days, with the order randomized across participants. The two sessions were separated by at least one day, but subjects were given up to a week after finishing Session 1 to complete Session 2. No video clips were repeated across the two sessions within the same participant.

Additional experiments

The procedures for Experiment 1 followed the general method described above. The remaining experiments were modified from this method, with the differences as follows.

Experiments 2 and 3: Feedback manipulations

In Experiment 2 (no feedback), instead of receiving feedback regarding accuracy on every trial, observers were no longer given any trial-wise feedback. The experiment proceeded to the next trial following a 500-ms gray screen. In Experiment 3 (partial feedback), instead of receiving feedback regarding accuracy on every trial, observers only received feedback for missed hazards. In other words, if they responded “hazard absent” on a hazard-present trial, they were shown a screen indicating that they missed a hazard.

In both Experiments 2 and 3, as in Experiment 1, participants were still shown a screen every 60 trials, indicating the total proportion correct in all trials up to that point.

All other aspects of the main experiment were the same in Experiments 2 and 3.

Experiment 4: Response correction

After each trial, observers were given the opportunity to correct their response. After pressing the up or down arrow key (to indicate a hazard-absent or hazard-present trial, respectively), observers were asked either to (1) press the space bar to continue, or (2) press “x” to change their previous response. This was intended to give observers the opportunity to correct any potential motor errors. For consistency in our data quality measure across experiments, observers were not given the opportunity to correct their responses on catch trials. Observers received feedback regarding their accuracy on each trial (following any corrections if applicable).

Experiment 5: Lower prevalence and participant prebriefing

Each participant completed two prevalence conditions, with lower prevalence levels than used previously (10% vs. 1% instead of 50% vs. 4%). The 10% condition consisted of 400 road videos (40 containing a hazard, 360 without a hazard), and the 1% condition consisted of 500 road videos (five containing a hazard, 495 without a hazard). As before, each of these conditions included an additional 40 catch trials involving an indoor/outdoor scene discrimination task, and all participants completed a 40-trial practice at 50% prevalence before starting each session of the experiment. The total number of trials in the 10% and 1% sessions (including the catch trials and practice) were 480 and 580, respectively.

In addition, we included a prebriefing manipulation to vary participants’ knowledge of the prevalence effect. Participants were randomly assigned to one of two groups (with 16 participants each): one group was not given any information about the nature of the prevalence effect (identical to our previous experiments), and the other group was informed about the prevalence effect and instructed to avoid it. The latter group was shown instruction screens (in between the practice and beginning the experiment) in each session with the following text: “Research has shown that people often miss, or fail to notice, rare events (like a hazard on the road). It is important that you look for the rare hazards among these videos. Remember: if you don’t find it often, you often don’t find it.” As in the previous experiments, both groups in the 1% and 10% conditions were informed that “unlike the practice, hazards will be relatively rare.”

Analysis

Each observer’s data were assessed for overall quality using two preregistered inclusion criteria. First, observers

needed to perform the catch trials (the indoor/outdoor scene discrimination task) at 85% accuracy or higher (note that chance is 50% in a two alternative forced choice task). Second, data from any observers who did not complete both sessions were discarded. These observers were replaced until we had a complete set of data (16 participants each in Experiments 1–4, and 32 in Experiment 5). Across all experiments, participants completed the catch trials with a mean accuracy of 98.4% ($SD = 2.1\%$), and no participants performed below the 85% minimum. One participant in Experiment 2 and one participant in Experiment 4 were replaced for not completing both sessions. In Experiment 5, one additional participant's data were removed from the analysis, and replaced with another participant for exceeding the maximum session duration allocated by Prolific (87 minutes for a 30-minute session). In total, three participants of 96 were replaced across the five experiments.

For each participant and prevalence condition, we calculated the miss rate as the proportion of hazard-absent responses out of the total number of hazard-present trials. The false alarm rate was equal to the proportion of hazard-present responses out of the total number of hazard-absent trials. We additionally calculated observers' sensitivity (d') and criterion in each condition using a signal detection analysis. To handle extreme values (e.g., hit or false-alarm rates of 0% or 100%), we calculated adjusted hit and false-alarm rates for all participants using the log-linear method described in (Hautus, 1995). Sensitivity (d') was then calculated as $\text{norminv}(\text{hit}_{\text{adj}}) - \text{norminv}(\text{false alarm}_{\text{adj}})$. Criterion (c) was calculated as $(\text{norminv}(\text{hit}_{\text{adj}}) + \text{norminv}(\text{false alarm}_{\text{adj}}))/-2$. Positive criterion values indicate a larger proportion of hazard-absent responses, negative values indicate a larger proportion of hazard-present responses, and values near zero indicate an equal proportion of each.

In Experiments 1–4, participants' error rates were analyzed using separate 2 (prevalence: low vs. high) \times 2 (error type: misses vs. false alarms) \times 2 (session order: low prevalence first or second) mixed-model ANOVAs. Prevalence and error type were within-subjects factors, and session order was a between-subjects factor. Effects of prevalence on hits and false alarms were followed up separately with pairwise contrasts. Sensitivity (d') and criterion (c) were analyzed with 2 (prevalence condition) \times 2 (session order) mixed-model ANOVAs. For Experiment 5, we included prebriefing condition (with or without) as a between-subjects factor. We note that the inclusion of session order as a between-subjects factor in these models was not preregistered but was added to account for any possible order effects on participants' performance. However, we note that session order did not significantly modulate the prevalence effect in most experiments (for Experiments 1–4, the three-way (Order \times Prevalence \times Error Type) interactions on error rate were all non-significant, $F(1, 14) < 2.2, p > .16$, and there

were no significant two-way Order \times Prevalence effects on criterion, $F(1, 14) < 3.9, p > .07$. In Experiment 5, there was a significant interaction between session order and criterion, $F(1, 28) = 11.88, p = .002$, indicating that participants who completed the 10% prevalence session first had a larger criterion shift (i.e., larger prevalence effect) between the 1% and 10% prevalence conditions (criterion difference of 0.63 vs. 0.26).

Results

Experiment 1

In Experiment 1, we observed significant main effects of both prevalence, $F(1, 14) = 9.27, p = .009, \eta_p^2 = .40$, and error type, $F(1, 14) = 65.5, p < .001, \eta_p^2 = .82$, which were qualified by a Prevalence \times Error Type interaction, $F(1, 14) = 44.75, p < .001, \eta_p^2 = .76$, consistent with the low-prevalence effect. When hazardous events were rare (4%), participants missed 39.7% of them, more than twice the rate observed in the high prevalence (50%) condition (18.2%), $t(14) = 4.94, p < .001$ (Fig. 2a). Conversely, the false-alarm rate was reduced at low prevalence (1.4% vs. 8.6%), $t(14) = 9.46, p < .001$ (Fig. 2b). Consistent with prior findings (J. M. Wolfe et al., 2007), response criterion became more conservative (more hazard-absent responses: $c = 0.99$ vs. $c = 0.24$), $t(14) = 11.1, p < .001$ (Fig. 2c), and d' did not change significantly ($d' = 2.53$ vs. $d' = 2.31$), $t(14) = 1.83, p = .09$ (Fig. 2d).

One possibility is that the low-prevalence effect could be an artifact of different numbers of target-present trials between conditions. At the very least, using only 20 hazard-present trials at low prevalence is likely to increase the variability in error rates, relative to the high-prevalence condition, which had 220 hazard-present trials. More problematically, participants may also be less familiar with hazard present trials in the low-prevalence condition, and this might induce some systematic bias. To address these possibilities, we separately analyzed only the first 40 trials that the participants saw in the 50% prevalence condition. These 40 trials contain 20 hazards on average. As shown in Fig. S1 in the Supplemental Information, we observed a similar low-prevalence effect in Experiment 1 (39.7% vs. 18.9%) and, indeed, across all of our experiments after equating the number of hazard-present videos between conditions. This result aligns with similar analyses in visual search (J. M. Wolfe et al., 2005), and indicates that our results are unlikely to be an artifact of differences in the number of target-present trials.

Another analytic consideration is that the relatively low proportion of false alarms at low prevalence could, in principle, produce unreliable estimates of d' due to a floor effect (Thomson et al., 2016). However, we attribute the

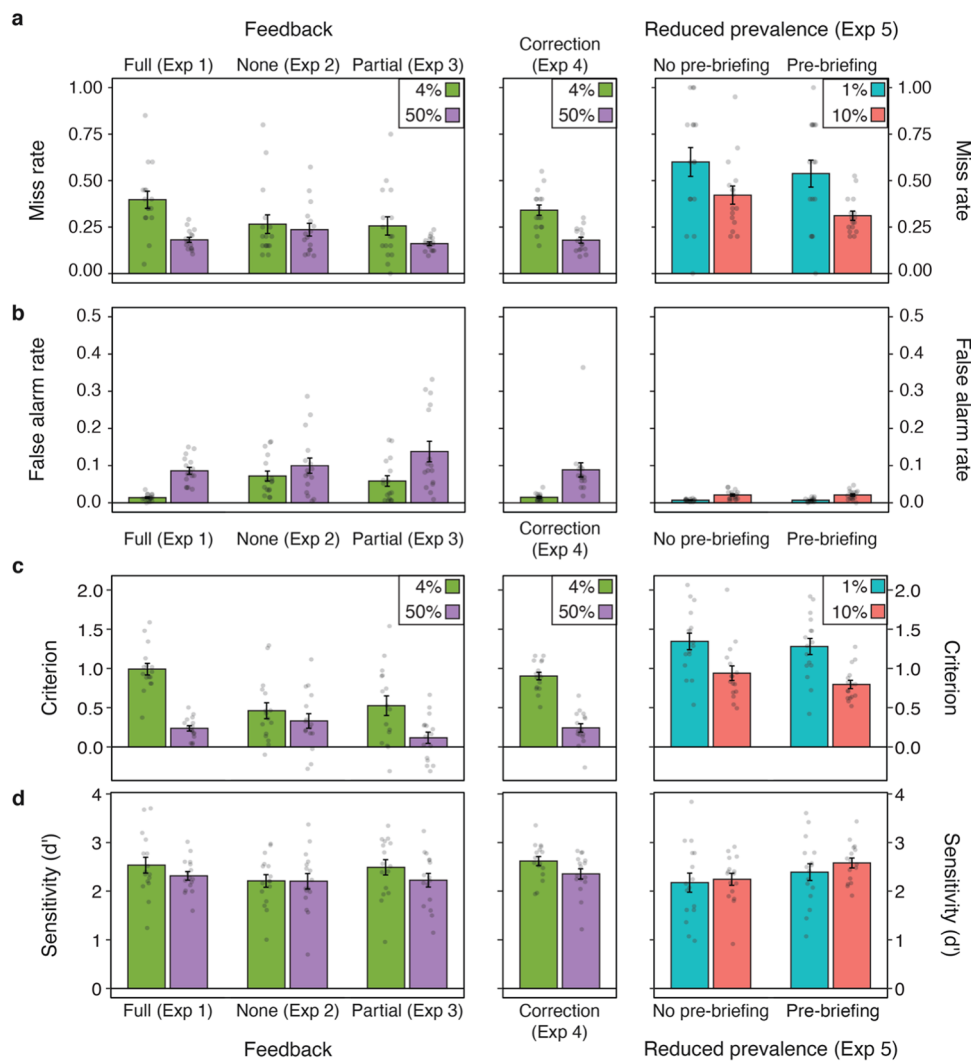


Fig. 2 Miss rates, false-alarm rates, criterion, and sensitivity for Experiments 1–5. **a** Miss rates in each experiment. In Experiments 1–3 (left panel), we varied the feedback provided to participants. When feedback was provided on each trial (Experiment 1), observers missed a significantly greater proportion of hazards under low prevalence (green) compared with high prevalence (purple). This effect was eliminated when no trial-wise feedback was provided (Experiment 2) and reduced when feedback was only provided for missed hazards (Experiment 3). In addition, miss rates were significantly higher under low (compared with high) prevalence when participants were given the opportunity to correct any potential motor errors (“finger errors”; Experiment 4). Finally, miss rates were significantly higher in Experiment 5, with 1% (teal) compared

with 10% (coral) prevalence, with no interaction between prebriefing conditions. **b** False-alarm rates were significantly higher under high prevalence compared with low prevalence in each experiment, except, as expected, for Experiment 2 (in which no feedback was provided). **c** Consistent with the pattern of miss and false-alarm rates, we observed significantly higher criterion values in each experiment under the lower prevalence condition compared with the higher prevalence condition, except for Experiment 2, in which no trial-wise feedback was provided. **d** Sensitivity (d') was similar, and not significantly different between the low- and high-prevalence conditions in most experiments (for Experiments 1, 2, 5), though it was significantly higher under low prevalence in Experiments 3 and 4. (Color figure online)

increase in misses at low prevalence to a criterion shift, rather than a decrease in d' , for two reasons. First, given the large number of trials used to estimate false alarm rates, this criterion shift was highly consistent across participants—at low prevalence, we observed an increase in miss rate for 14 of 16 participants and a drop in false-alarm rates in all 16 participants. Moreover, the observed decrease in hit rate at low prevalence (z -transformed

values for high vs. low: 0.90 vs. 0.25, a difference of 0.65) was accompanied by a commensurate decrease in false-alarm rate (-1.36 vs. -2.17 , a difference of 0.81). Second, if we did have a floor effect in false alarm rates, this would produce an *underestimate* of d' as prevalence decreases. Given that d' is similar across our two prevalence conditions, a correction to any underestimate would produce higher (not lower) d' values at low prevalence.

Our study was diverse in the demographics of its participants and its images. Participants were from the U.S., Canada, and the UK. The stimuli included both left-sided and right-sided driving (14.6% and 85.4% of trials, respectively). One might imagine that the match between participant and stimulus origin could be important. Prior work has suggested that familiarity may impact drivers' ability to detect changes in their environment, ranging from hazards (Thompson & Sabik, 2018) to changes in road signs (Martens, 2018; Martens & Fox, 2007a, 2007b), although these effects may require deep familiarity with specific road segments, rather than the larger context (Beanland & Wynne, 2019; Wynne et al., 2019). Thompson and Sabik's results, demonstrating impaired performance when drivers view road scenes with an unfamiliar traffic direction (left vs. right), suggest that our effects could be seen inconsistently across different demographics, or that they could be restricted to a subset of participants or videos.

To test for this, we separately analyzed participants' responses at 4% and 50% prevalence (across Experiments 1 and 4), based on country of residence and traffic direction in the video (left or right-handed traffic). These results are shown in Figs. S2 and S3 in the Supplemental Information. Briefly, we observed a consistent low-prevalence effect (higher miss rates under the low-prevalence condition compared with the high-prevalence condition) in both North American and UK participants (UK: 37.6% vs. 17.4%), $t(29) = 6.93$, $p < .001$; North America: 35.0% vs. 20.3%), $t(29) = 3.48$, $p = .002$, with no difference in the size of this effect across participant groups (i.e., no Country \times Prevalence \times Error Type interaction), $F(1, 29) = .12$, $p = .73$. Furthermore, we observed higher miss rates in the low prevalence condition compared with the high prevalence condition when we restricted our analysis to cases where the participant location of residence matched the the traffic direction in the video (i.e., North American participants and videos with right-handed traffic; miss rates of 34.3% vs. 20.9%), $t(29) = 2.91$, $p = .007$. Therefore, the effect we observe seems broadly consistent across participant demographics and is not restricted to a single subset of our sample. Comparing these error rates across different populations in a larger sample of participants or in a more naturalistic setting would be an important area for future follow-up work.

Experiments 2 and 3: Feedback manipulations

In Experiments 2 and 3, we tested the effect of feedback on the observed criterion shift. Previous work has shown that the LPE may be modulated by trial-wise feedback, such that observers adopt a more conservative criterion with feedback, and a more liberal criterion without feedback (Lyu et al., 2021). In Experiment 2, a new set of 16 participants performed the same task as

in Experiment 1 but received no feedback on their responses. Here, we no longer observed a significant prevalence effect on error rate (Prevalence \times Error Type interaction), $F(1, 14) = 0.92$, $p = .35$, $\eta_p^2 = .06$, with a comparable proportion of missed hazards under low versus high prevalence (26.6% vs. 23.6%, respectively), $t(14) = 0.85$, $p = 0.41$ (Fig. 2a). Similarly, there was no significant difference in d' (low: 2.21 vs. high: 2.20), $t(14) = 0.30$, $p = .77$, or criterion (low: 0.46 vs. high: 0.33), $t(14) = 0.98$, $p = .34$, between the low- and high-prevalence conditions (Fig. 2c–d).

In Experiment 3, we provided feedback only when participants failed to report a hazard. Here, the LPE was somewhat reduced, with a significant Prevalence \times Error Type interaction, $F(1,14) = 10.15$, $p = .007$, $\eta_p^2 = .42$, and a smaller difference in proportion of missed hazards (low prevalence: 25.6% misses vs high: 16.1%), $t(14) = 2.24$, $p = .042$ (Fig. 2a). At low prevalence, we also observed a more conservative criterion (low: 0.52 vs. high: 0.12), $t(14) = 3.88$, $p = .002$, along with a small but significant increase in d' (low: 2.49 vs. high: 2.22), $t(14) = 2.22$, $p = .044$.

Experiment 4: Response correction

Experiment 4 addresses the question of whether elevated miss rates at low prevalence can be attributed to motor errors (Fleck & Mitroff, 2007). In other words, are participants simply pushing the “no hazard” button out of habit? To determine whether this was the case, we gave participants the opportunity to correct their errors on each trial. We continued to see a significant increase in the miss proportion at low prevalence. We observed a significant Prevalence \times Error Type interaction, $F(1, 14) = 41.86$, $p < .001$, $\eta_p^2 = .75$, with an elevated miss rate at low prevalence (low: 34.1% missed hazards vs. high: 18.0%), $t(14) = 6.52$, $p < .001$ (Fig. 2a). Error rates were consistent with a more conservative criterion (low: 0.90 vs. 0.24), $t(14) = 8.85$, $p < .001$, at low prevalence. This was accompanied by a smaller, but significant increase in d' in the low prevalence condition (2.62 vs. 2.35), $t(14) = 2.18$, $p = .047$.

Experiment 5: Lower prevalence and participant prebriefing

While Experiments 1–4 suggest that infrequent hazards are missed, there are still remaining questions with regard to the real-world relevance of these findings. First, one might ask whether miss rates would continue to increase if we reduced hazard prevalence further. This would be expected from prior research with static stimuli (Mitroff & Biggs, 2014) and more closely approximate the real prevalence of near-collision situations on the road. The 50% prevalence rate, used as a point of comparison in Experiment 1, may be simply too large and not at all representative of hazard

prevalence in *any* road environment since, outside of a war zone or a *Jurassic Park* film, hazards are simply not that frequent. Investigating differences between different low prevalence rates (e.g., between 10% and 1%) may be more informative. Second, can the LPE be reduced or eliminated by informing participants of this effect, and emphasizing the importance of finding the rare hazards?

In Experiment 5 ($n = 32$), we reduced prevalence to 1% and 10%, and assigned participants to one of two groups ($n = 16$ each). One group was informed of the prevalence effect and instructed to avoid it, and the other was not. As before, we observed a significant Prevalence \times Error Type interaction, $F(1, 28) = 28.63$, $p < .001$, $\eta_p^2 = .51$, with a significantly larger proportion of missed hazards at 1% prevalence compared with 10% prevalence (56.9% vs. 36.6%), $t(28) = 4.99$, $p < .001$. However, neither the main effect, $F(1, 28) = 1.32$, $p = .26$, nor any of the interactions involving prebriefing condition, $F(1, 28) < 1.30$, $p > .26$, were significant, indicating that informing observers about the prevalence effect did not significantly affect it. As in the other experiments, one limitation of this design is that a single miss can have a very large effect on error rates, particularly in the lower prevalence condition, where there are only five hazard-present trials. To equate the number of hazard-present trials between conditions, for the 10% prevalence condition, we analyzed only the first 50 trials that participants saw (these contain five hazards on average). As in our other experiments, we see a similar pattern of results, with miss rates of 56.9% at low prevalence and 30.2% at high prevalence (Fig. S1).

Finally, we compared sensitivity and criterion across the low-prevalence conditions in our experiments. Figure 3 plots d' and criterion across rates of 4% and 50% from Experiment 1, and rates of 1% and 10% from Experiment 5. Consistent with the results from individual experiments, we observed a significantly more conservative criterion at 1% compared with 4% prevalence, $t(46) = 2.78$, $p = .008$, with no difference in d' , $t(46) = 1.16$, $p = .25$.

Discussion

Previous work has demonstrated a low-prevalence effect (LPE) in target detection and visual search tasks, in which observers frequently miss targets as they become increasingly rare (e.g., Broadbent & Gregory, 1965; Colquhoun, 1961; J. M. Wolfe et al., 2005). Here we reproduced this phenomenon in a novel hazard detection paradigm, in which observers monitored brief clips of road videos for hazards, extending the LPE literature to dynamic natural scenes. Miss errors were elevated at low prevalence despite the heterogeneity, complexity, and dynamic nature of these events and despite the importance of accurate response to such hazards

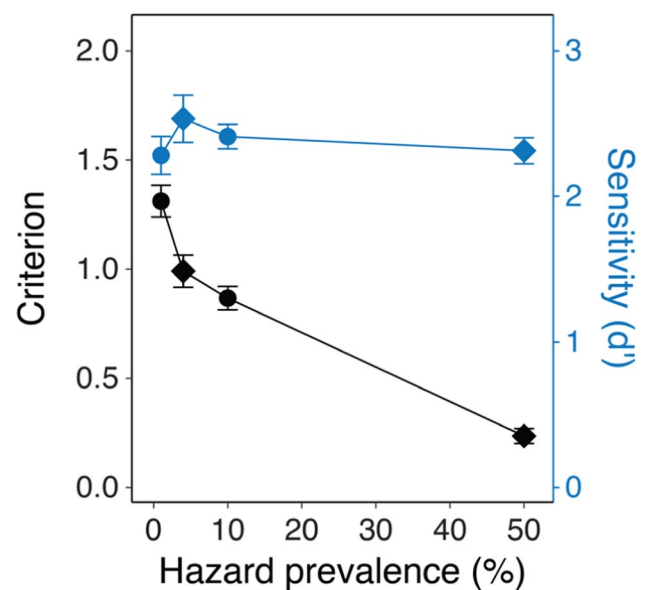


Fig. 3 Plot of criterion (left y-axis, black) and d' (right y-axis, blue) across different prevalence rates. Results from several experiments show that d' remains stable across prevalence, indicating no loss of sensitivity when hazards are rare. In contrast, criterion changes dramatically as hazards become less common. Higher criterion values reflect a greater tendency to respond “no hazard” and thus to miss more hazards. (Color figure online)

in the real world. In follow-up experiments, we demonstrated that this effect is dependent on receiving immediate feedback (Experiments 2 & 3), cannot be attributed to motor errors (Experiment 4), and scales with decreasing prevalence, while being resistant to participant prebriefing (Experiment 5). The low-prevalence effect was also highly consistent across our sample; as shown in Figure S4, in the experiments with full feedback (Exp 1, 4, and 5), 54 of 64 observers showed miss rates consistent with the LPE (higher proportion of misses under low prevalence).

Consistent with previous work, across our experiments, we observed a more conservative criterion (i.e., fewer “hazard present” responses) as hazard prevalence was reduced. Importantly, this criterion shift was seen consistently as prevalence was reduced from 50% down to 4% and 1% (Fig. 3), matching results reported previously in visual search, which show that increasingly rare targets (i.e., less than 5%) are even more likely to be missed (Mitroff & Biggs, 2014). In contrast, sensitivity (d') remained relatively stable across the prevalence rates tested. This result also aligns with work in vigilance (Thomson et al., 2016) that identifies a shift in criterion rather than sensitivity as a source of vigilance performance decrements. However, we note that, in two of our experiments (Experiment 3: partial feedback; Experiment 4: response correction) we additionally saw a small, but significant increase (rather than a decrease) in d' at low prevalence. This result has been observed previously

for the LPE in visual search (J. M. Wolfe et al., 2007) where it may reflect unequal variance in the underlying signal and noise distributions, which renders interpretation of d' more difficult. Nevertheless, as shown in Fig. 3, any LPE change in d' is small compared with the change in criterion. In other words, consistent with other tasks, as hazard prevalence decreases, criterion becomes more conservative and miss errors become more frequent.

In addition, we reproduce recent work demonstrating that the LPE may depend on trial-wise feedback. Lyu et al. (2021) showed that, in categorizing ambiguous stimuli (e.g., colored patches, shapes), observers adopt a more conservative criterion at low prevalence when provided with feedback, while adopting a more liberal criterion without feedback. Although we did not see a more liberal criterion at low prevalence (4%) compared with high prevalence (50%) without feedback (Experiment 2), this manipulation was sufficient to eliminate the LPE. As expected, we observed the largest criterion shift when observers received feedback regarding response accuracy after each trial. Translating these findings from experimental contexts to the real world has a number of limitations, but it is likely that much of the feedback that drivers receive is self-generated (e.g., “I did not need to brake”) or dramatic (e.g., a literal “hit”). Many real-world errors will be false alarms (no need to brake). Self-generated feedback about all those false alarms may make observers less likely to respond to actual hazards.

Driving can be considered a vigilance task. However, our results differ from some findings in the vigilance literature. Target heterogeneity can attenuate the vigilance decrement (Thiffault & Bergeron, 2003; Thomson et al., 2015), but we still see a robust effect of hazard prevalence using a heterogeneous set of real road hazards (e.g., pedestrians, animals, unconstrained objects, other vehicles). This suggests that target complexity and variability are insufficient to prevent the LPE. Moreover, unlike some vigilance tasks, where increased effort, or simply telling participants to try harder (N. H. Mackworth, 1950) may attenuate the effect, our LPE is not easily cognitively penetrable. When informed of the prevalence effect and instructed to avoid it, participants nevertheless missed a higher proportion of hazards in the lower prevalence condition in Experiment 5.

While we have established a clear low-prevalence effect when observers view the dynamic road scenes from our stimulus set, there are many differences from real-world driving that limit the generalizability of these findings. In particular, viewing videos of dynamic road scenes is inherently a much more passive task compared with operating a vehicle, which requires constant monitoring, maneuvering, and correction. There is a lot less at stake when the participant is not actually behind the wheel. Future work would also be necessary to determine whether the effects we observe persist with longer video durations in more continuous settings (i.e., as opposed

to a series of brief clips), as well as how these effects may vary with the driver’s experience, skill, and familiarity with the road environment. Additionally, it would be useful to have a larger sample of participants to examine cultural and other demographic variables. This is particularly important, given that the frequency of different types of hazards (e.g., pedestrians vs. vehicles) varies considerably across different driving environments.

Another important difference from real-world driving are the costs and payoffs. Could the high costs for missed hazards be sufficient to eliminate this low-prevalence effect in the real world? In principle, decision theory would predict that changing the payoffs could minimize or reduce this criterion shift. However, previous work has indicated that payoffs may be less effective than changes in target prevalence in shifting observers’ criterion (Healy & Kubovy, 1981; Maddox, 2002). Moreover, the high real-world costs for missed targets appear to be insufficient to eliminate the LPE in some visual search tasks, where experts miss important targets like abnormalities in medical images (Evans et al., 2013). Determining whether this is also the case for on-road hazard detection would require further work, and a closer focus on driver expertise and environment (e.g., developing a stimulus set for one environment and recruiting observers from that same environment).

Indications of an LPE are present elsewhere in driving research. For example, others have reported a “safety in numbers” effect, in which higher prevalence of pedestrians and cyclists actually reduces the likelihood that they will collide with a motor vehicle (Beanland et al., 2015; Jacobsen, 2015). Similarly, the driving literature points to the dangers of monotonous driving (Thiffault & Bergeron, 2003), where drivers must respond infrequently. These effects could be exacerbated by drivers’ assumptions about particular settings. A busy street may keep a driver on their toes, but the same driver may fail to respond in time to an animal that unexpectedly jumps into a lonely desert road. Finally, and somewhat perversely, prevalence effects may become more of a problem as vehicle automation makes vehicles and roads safer. Others have drawn attention to driver “deskilling” as an unintended consequence of automation’s ability to remove many dangerous situations from the road (Noy et al., 2018). A semiautonomous vehicle, where the driver may be expected to act *only* when a hazard arises, may create exactly the type of low-prevalence situation in which humans perform badly, although whether takeover situations will exist, or if drivers will notice such situations on their own is an open question (de Winter et al., 2021). Determining how to improve drivers’ ability to respond to the few hazardous situations that remain will be an important question for future work.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13423-022-02159-0>.

Acknowledgments This work was supported by NSERC (Discovery Grants RGPIN-2021-02730 to B.W. and RGPIN-2022-03131 to A.K.). We thank Simran Kanda and Silvia Guidi for assistance with data collection and video processing, and Ben Sawyer for helpful discussions.

Author contributions A.K., J.M.W., and B.W. conceived and designed the experiments. A.K. and B.W. performed the experiments and analyzed the data. A.K., J.M.W., and B.W. wrote the manuscript.

Open practices statement All data and materials, including stimulus and analysis scripts, are available on the Open Science Framework (OSF) online (<https://osf.io/r9uk7/>). Preregistration information for each experiment can be accessed online (<https://osf.io/r9uk7/registrations>).

References

- Baddeley, A. D., & Colquhoun, W. P. (1969). Signal probability and vigilance: A reappraisal of the “signal-rate” effect. *British Journal of Psychology*, *60*(2), 169–178.
- Beanland, V., Lenné, M. G., & Underwood, G. (2014). Safety in numbers: Target prevalence affects the detection of vehicles during simulated driving. *Attention, Perception, & Psychophysics*, *76*(3), 805–813.
- Beanland, V., Lenné, M., Underwood, G., & Rößger, L. (2015). Psychological factors in seeing motorcycles. In L. Rößger, M. G. Lenné, & G. Underwood (Eds.), *Increasing motorcycle conspicuity: Design and assessment of interventions to enhance rider safety* (pp. 21–50). Ashgate.
- Beanland, V., & Wynne, R. A. (2019). Does familiarity breed competence or contempt? Effects of driver experience, road type and familiarity on hazard perception. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 2006–2010.
- Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*(1), e9414–e9429.
- Broadbent, D. E., & Gregory, M. (1965). Effects of noise and of signal rate upon vigilance analysed by means of decision theory. *Human Factors*, *7*(2), 155–162.
- Colquhoun, W. P. (1961). The effect of ‘unwanted’ signals on performance in a vigilance task. *Ergonomics*, *4*(1), 41–51.
- de Winter, J., Stanton, N., & Eisma, Y. B. (2021). Is the take-over paradigm a mere convenience? *Transportation Research Interdisciplinary Perspectives*, *10*, 100370.
- Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, *113*(10), 2636–2641.
- Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If you don’t find it often, you often don’t find it: Why some cancers are missed in breast cancer screening. *PLOS ONE*, *8*(5), Article e64366.
- Evans, K. K., Tambouret, R. H., Evered, A., Wilbur, D. C., & Wolfe, J. M. (2011). Prevalence of abnormalities influences cytologists’ error rates in screening for cervical cancer. *Archives of Pathology & Laboratory Medicine*, *135*(12), 1557–1560.
- Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science*, *18*(11), 943–947.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (pp. xi, 455). John Wiley.
- Green, M. (2000). “How long does it take to stop?” Methodological analysis of driver perception-brake times. *Transportation Human Factors*, *2*(3), 195–216.
- Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472.
- Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d' . *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51.
- Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 344–354.
- Horowitz, T. S. (2017). Prevalence in visual search: From the clinic to the lab and back again. *Japanese Psychological Research*, *59*(2), 65–108.
- Jacobsen, P. L. (2015). Safety in numbers: More walkers and bicyclists, safer walking and bicycling. *Injury Prevention*, *21*(4), 271–275.
- Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. *Science* <https://www.science.org/doi/abs/10.1126/science.aap8731>
- Lyu, W., Levari, D. E., Nartker, M. S., Little, D. S., & Wolfe, J. M. (2021). Feedback moderates the effect of prevalence on perceptual decisions. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-021-01956-3>
- Mackworth, J. (1970). *Vigilance and attention: A signal detection approach*. Penguin Books.
- Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*(1), 6–21.
- Mackworth, N. H. (1950). Researches on the measurement of human performance. *Researches on the Measurement of Human Performance*, *268* <https://www.cabdirect.org/cabdirect/abstract/19502703894>
- Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*, *78*(3), 567–595.
- Martens, M. H. (2018). The failure to respond to changes in the road environment: Does road familiarity play a role? *Transportation Research Part F: Traffic Psychology and Behaviour*, *57*, 23–35.
- Martens, M. H., & Fox, M. (2007a). Does road familiarity change eye fixations? A comparison between watching a video and real driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(1), 33–47.
- Martens, M. H., & Fox, M. R. J. (2007b). Do familiarity and expectations change perception? Drivers’ glances and response to changes. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(6), 476–492.
- Mitroff, S. R., & Biggs, A. T. (2014). The ultra-rare-item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science*, *25*(1), 284–289.
- Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, Y., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., & Oliva, A. (2019). Moments in Time Dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1–1*. <https://doi.org/10.1109/tpami.2019.2901464>
- Noy, I. Y., Shinar, D., & Horrey, W. J. (2018). Automated driving: Safety blind spots. *Safety Science*, *102*, 68–78.
- Oliva, A. (2005). Gist of the scene. In L. Itti, J. Tsotsos, & G. Rees (Eds.), *Neurobiology of attention* (pp. 251–256). Elsevier.
- Parasuraman, R., & Davies, D. R. (1977). A taxonomic analysis of vigilance performance. In R. R. Mackie (Ed.), *Vigilance: Theory, operational performance, and physiological correlates* (pp. 559–574). Springer. https://doi.org/10.1007/978-1-4684-2529-1_26
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). *PsychoPy2: Experiments in behavior made easy* (pp. 1–9). <https://doi.org/10.3758/s13428-018-01193-y>

- Thiffault, P., & Bergeron, J. (2003). Monotony of road environment and driver fatigue: A simulator study. *Accident Analysis & Prevention*, 35(3), 381–391.
- Thompson, C., & Sabik, M. (2018). Allocation of attention in familiar and unfamiliar traffic scenarios. *Transportation Research Part F: Traffic Psychology and Behaviour*, 55, 188–198.
- Thomson, D. R., Besner, D., & Smilek, D. (2016). A critical examination of the evidence for sensitivity loss in modern vigilance tasks. *Psychological Review*, 123(1), 70–83.
- Thomson, D. R., Smilek, D., & Besner, D. (2015). Reducing the vigilance decrement: The effects of perceptual variability. *Consciousness and Cognition*, 33, 386–397.
- Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, 50(3), 433–441.
- Wolfe, B., Seppelt, B., Mehler, B., Reimer, B., & Rosenholtz, R. (2020). Rapid holistic perception and evasion of road hazards. *Journal of Experimental Psychology: General*, 149(3), 490–500.
- Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision*, 13(3), 33–33.
- Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, 435(7041), 439–440.
- Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*, 136(4), 623–638.
- Wynne, R. A., Beanland, V., Read, G. J. M., & Salmon, P. M. (2019). You look familiar: Learner driver hazard identification on familiar and unfamiliar roads. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), 2001–2005.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.