Behavior Research Methods
https://doi.org/10.3758/s13428-023-02299-8

ORIGINAL MANUSCRIPT

Road Hazard Stimuli: Annotated naturalistic road videos for studying hazard detection and scene perception

Jiali Song[1] · **Anna Kosovicheva**[1] · **Benjamin Wolfe**[1]

Accepted: 17 November 2023

Abstract

Driving requires vision, yet there is little empirical data about how vision and cognition support safe driving. It is difficult to study perception during natural driving because the experimental rigor required would be dangerous and unethical to implement on the road. The driving environment is complex, dynamic, and immensely variable, making it extremely challenging to accurately replicate in simulation. Our proposed solution is to study vision using stimuli which reflect this inherent complexity by using footage of real driving situations. To this end, we curated a set of 750 crowd-sourced video clips (434 hazard and 316 no-hazard clips), which have been spatially, temporally, and categorically annotated. These annotations describe where the hazard appears, what it is, and when it occurs. In addition, perceived dangerousness changes from moment to moment and is not a simple binary detection judgement. To capture this more granular aspect of our stimuli, we asked 48 observers to rate the perceived hazardousness of 1356 brief video clips taken from these 750 source clips on a continuous scale. These ratings span the entire scale, have high interrater agreement, and are robust to driving history. This novel stimulus set is not only useful for understanding drivers' ability to detect hazards, but is also a tool for studying dynamic scene perception and other aspects of visual function. While this stimulus set was originally designed for behavioral studies, researchers interested in other areas such as traffic safety or computer vision may also find this dataset a useful resource.

Keywords Driving · Road video dataset · Road hazards · Dashcam videos

Introduction

Vision is enormously important in daily life. It is the predominant sense we rely on to navigate and interact with the world, and is essential for tasks as mundane as deciding where to put our foot next, or as consequential as avoiding a moose running down the road while driving. Yet we tend to study vision using stimuli that seldom look like the world or the activities we engage in every day. Driving is an example of a real-world scene perception task we do all the time. Between 2019 and 2020, Americans spent on average 59 min a day driving (AAA Foundation for Traffic Safety, 2021). Safe driving requires early detection and avoidance of hazards and recognizing and obeying traffic signals (Green, 2000), and vision is the primary sensory modality that provides information needed to respond appropriately (Sivak, 1996). For example, a driver would only stop at a stop sign if they see the stop sign before they must begin braking, as no other information is available to the driver to indicate that they need to do so. The importance of vision in driving is also reflected in licensing criteria, as minimum visual acuity and field of view are common requirements for licensure internationally (Yan et al., 2019). It is clear and widely acknowledged that vision is incredibly important for driving (Schieber et al., 2009; Sivak, 1996).

There are multiple benefits of studying vision and driving at both applied and fundamental levels. Collisions are rare relative to the total number of driving hours, but they have a disproportionate societal impact. In 2020 alone, there were 38,824 fatalities due to traffic collisions in the United States (AAA Foundation for Traffic Safety, 2021). Studying the contributions of vision to driving can help us understand why collisions happen and reduce them by informing how we design driving-relevant technologies to maximize road safety for all road users. There are many efforts to reduce collisions, but much of the research in this space deals with infrastructural or technological solutions, such as more

✉ Jiali Song
jiali.song@utoronto.ca

[1] Department of Psychology, University of Toronto Mississauga, Mississauga, Canada

Published online: 11 December 2023

intuitive road design (Theeuwes, 2021), or cars that react for the driver (SAE International, 2018). The human aspect of the issue is critically important but comparatively less studied, particularly in perceptual contexts. To effectively address human error on the road, it is crucial to understand the cognitive and perceptual limitations of the human visual system that impede hazard detection.

Moreover, understanding the contributions of vision and cognition in driving can provide a more complete understanding of visual perception at a fundamental level, and necessitates understanding how they support daily function, such as driving. Yet we tend to study vision using stimuli that seldom look like the world or the activities we engage in outside of the laboratory. Unlike typical laboratory tasks of scene perception, driving is dynamic, consequential, and time-limited. Situations on the road can become hazardous in a split second, and road environments are situationally diverse and challenging to fully replicate in simulation. That said, we should note that in this work, we focused on overt, abrupt hazards rather than latent hazards, for the simple reason that they are less ambiguous to code and provided us with a more broadly useful dataset for perceptual research. While latent, potential hazards are themselves perceptually interesting, and very relevant to road safety, they were not the focus of this effort.

Critically, the immense time pressure and high stakes experienced by drivers on the road is difficult (and unethical) to replicate in the lab. Hazards can appear suddenly, and drivers can detect the hazard, decide on the best corrective measure, and enact it within as little as 1.5 s (Green, 2000), while the hazard is unfolding. However, brake reaction time can be as short as 0.7 s when drivers are able to anticipate a hazard (Green, 2000), suggesting that the ability to predict the onset of hazards has significant impact on reaction time. However, laboratory studies suggest that observers can make global judgements about the gist of a static scene in as little as 100 ms (Li et al., 2002; Navon, 1977; Oliva, 2005). Presumably, drivers are also able to make similar global decisions about driving scenes, but the dynamic environment complicates this inference. Although static visual images contain some information drivers need to detect hazards, moving scenes provide more information (e.g., travel speed and direction) which improves scene processing and hazard detection (Moharrer et al., 2020; Wolfe et al., 2019, 2020). Furthermore, given the role that anticipating hazards has on a driver's ability to respond quickly, there is likely information in a road scene prior to when a hazard has appeared that may contribute to the driver's decision even before the hazard has appeared on the scene. However, it is unclear which specific aspects of the road scene drivers use to make these predictive inferences. For these reasons, it is essential to study dynamic scenes in addition to static ones to gain a full appreciation of how drivers cope with these demands.

Furthermore, road situations are diverse and difficult to replicate in simulation. Some hazards are not very dangerous, like a paper bag floating down the road, other hazards, like a truck overturning ahead, are extremely dangerous. This adds another level of complexity that cannot be captured by simple categorization tasks commonly used in scene perception studies. While methods like staged hazards in controlled driving (Falkmer & Gregersen, 2001; Mourant & Rockwell, 1972; Underwood et al., 2002) or in driving simulators (Beanland et al., 2014; Duivenvoorden et al., 2015; Schall et al., 2013) can capture some hazards, they are unlikely to capture the range of hazards that drivers can encounter on the road in the real world. For example, controlled driving studies cannot endanger participants, thereby constraining the types of hazards that can be studied. Hazards can be emulated safely in the simulator, but the reduction in visual fidelity and reliance on computer physics models mean that simulator scenarios could look quite different from real hazards. Moreover, the types of hazards that can be used in these laboratory scenarios are limited by the imagination of the researcher, excluding many hazards that drivers may encounter on the road from study. Researchers need to account for such variability in road hazards to fully understand how the visual system deals with these demands.

Given that it is impossible and unethical to replicate on-road conditions inside the lab, one solution is to bring the road into the lab through a richer set of stimuli. Although there are several publicly available road video datasets (Chan et al., 2017; Geiger et al., 2013; Saunier et al., 2014), they are unsuitable for perception studies be because the footage was gathered for computer vision, rather than for human observers. For example, many datasets are gathered by cameras outside of the vehicle from a perspective that is virtually never experienced by drivers (Geiger et al., 2013; Saunier et al., 2014). Such footage is unlikely to elicit behaviors comparable to natural driving because they deviate significantly from a driver's visual input.

Moreover, hazardous situations on the road have different demands from safe scenarios (Crundall et al., 1999). Datasets for studying human perception need to include both hazardous and safe situations. However, the rarity of hazards makes it difficult to obtain them from naturalistic driving studies. For example, in the CanDrive/OzcanDrive study (Marshall, Man-Son-Hing et al., 2013a, Marshall, Wilson et al., 2013b), only 139 out of 1207 drivers (or 11%) experienced one or more crashes (Langford et al., 2013). Another example, SHRP2 (Campbell, 2012) recorded over 3500 crashes and near crashes, extracted from over 4300 years of naturalistic driving data (Hankey et al., 2016). While it is certainly possible to extract collisions and near-collision events from datasets like SHRP2, this required not only tens of thousands of hours of recorded road video, but also the equipment, personnel, time, and funds to do. While these studies provide an invaluable window into driver behavior, they were never designed to provide stimuli for perceptual studies of what the driver sees, or of dynamic scene perception more generally. In addition, the data in these large-scale naturalistic studies are often geographically confined, which makes it difficult for researchers to study perception in new, unfamiliar environments. To address our need for a stimulus set comprised of real driving situations, focused on dangerous events, we have developed an annotated video dataset for studying road scene perception, the *Road Hazard Stimuli*.

Road hazard stimuli

We developed the Road Hazard Stimuli to address these challenges to studying drivers' perception of dynamic natural road scenes. We collected and annotated a set of 750 video clips of real on-road situations, taken from dashcam videos shared on social media websites YouTube and Reddit. The set consists of 434 video clips containing hazards, and 316 control video clips extracted from hazard-free portions of the longer video segments corresponding to these hazards. We defined hazards as events in the video that required an immediate evasive maneuver from the camera vehicle. Given that a driver typically encounters a collision once every 10 years, this dataset represents over 4300 years of cumulative driving experience (using the method of Horswill et al., 2021). It has potential to aid research not only in traffic safety and driver behavior but also as a tool for a wider range of perceptual and computational studies.

The footage was recorded from front-facing dashcams, near the driver's perspective inside the vehicle. The footage captures most of what a driver would see ahead of their vehicle while preserving the temporal characteristics of hazards. Observers are then able to view the scene naturally as they would from inside a car in a real on-road scenario. This similarity is crucial because it is more likely to elicit behavior that is comparable to real driving, and provides a starting point for understanding how drivers extract information from road environments.

This video set was based on a previous dataset (Wolfe et al., 2019), and significantly expands it by nearly doubling the number of hazard videos. To make the videos maximally useful for the largest number of behavioral researchers, we spatially and temporally annotated each video, and categorized it based on several features (see Table 1 for categorical annotations for all videos, and Table 2 for categorial annotations specific to hazard videos).

In addition, we gathered descriptive data about the hazardousness of the videos in the dataset, and to examine the extent to which perceived hazardousness of these videos is binary or continuous. To this end, we asked 48 licensed drivers to rate the hazardousness of 1356 brief video clips (432 hazard video clips and 924 no-hazard video clips extracted from the original set of 750 videos at least 8 s in duration). As hazardous road situations unfold gradually over time, we asked participants to rate the hazard-present videos at the same critical moment in each video, during the time window immediately preceding the driver's response. The duration of the excerpts was chosen to mimic the time pressure of real driving. To ensure participants are still able to detect hazards, we chose a duration of 333 ms, which is 150% of the threshold duration that young adult drivers needed to accurately detect a hazard, based on a previous study using a subset of the Road Hazard Stimuli (Wolfe et al., 2020).

These data provide more granular information about the variability of hazards in the dataset, which is an important aspect of understanding hazards that cannot be captured by simple binary classification. Given that driving experience may modulate hazard detection ability (Borowsky et al., 2008; Cooper et al., 1995; Jackson et al., 2009), we also asked participants to complete a driving history survey,

Table 1 Categorical annotations for all videos in Road Hazard Stimuli

Category	Hazard videos ($n = 434$)	No hazard videos ($n = 316$)	Total ($n = 750$)
Traffic convention			
Left-hand traffic	74	42	116
Right-hand traffic	360	274	634
Time of day			
Day	374	284	658
Night	60	32	92

Table 2 Categorical annotations for hazard videos

Category	No. of hazard videos ($n = 434$)
First driver response	
Brake	352
Swerve left	44
Swerve right	40
Hazard type	
Animal	25
Cyclist	6
Obstacle	9
Pedestrian	29
Vehicle	357
Other	10
Hazard location	
Screen left	152
Screen right	134
Crosses both sides	150

which allowed us to examine the relation between hazardousness ratings and driving experience. In this manuscript, we describe the contents of the dataset, how it was created, and its potential uses.

Methods

Video sources

To find videos that best capture a variety of real hazards drivers may encounter on the road, we took advantage of the proliferation of dashcams. Dashcams are inexpensive and widely available, and drivers use them to record footage of their environments during driving for insurance purposes. Some drivers also share their dashcam footage on social media so other people can view and comment on them. A useful side benefit of these videos is that they provide a corpus of rare, real-world events genuinely encountered during driving. Moreover, the dashcams are placed inside the car, from a perspective near the driver, making them well suited for studying how drivers perceive the road environment.

Inclusion criteria

There were several inclusion criteria for videos. Videos needed to be recorded from a front-facing dashcam from inside the vehicle. We excluded footage from backward-facing cameras, cyclists, and motorcyclists, since these would produce a different view of the road than our focus. Videos were required to include a distinct hazard that the driver needed to react to immediately to avoid a collision. In addition, the driver's reaction to the hazard needed to be visible in the footage (e.g., slowing, indicative of a braking response, or a shift in camera viewpoint, indicating a maneuver to the left or right). Videos were also required to be free from obstruction (watermarking, icons, annotations) such that the road situation is visible, and videos were required to be at least 10 s long with pixel resolutions of at least 720p, in landscape/wide-screen format.

Exclusion criteria

Videos in which drivers did not respond, and videos that contained severe personal injury or other graphic events were excluded. These criteria allowed for the inclusion of a wide array of different kinds of hazards, such as animals, pedestrians, vehicles, and other types of hazardous objects (Fig. 1), while minimizing the risk of emotional trauma to research participants and annotators.

Video processing

Using these criteria, we identified and downloaded 511 source videos from YouTube and Reddit. For all videos, we resized them to 720p, 30 fps video format, and removed all audio tracks to remove potential sources of distraction. Most source videos contained a hazardous event and were

Fig. 1 Examples to illustrate each hazard type. Each image is cropped from a single frame to illustrate the hazard. Videos in the stimulus set contain more complete road scenes than shown here. **A** A kangaroo jumps onto the road from the left. **B** A child on a bicycle enters the intersection from the left. **C** A white plastic cannister flies down the road towards the camera car. **D** A pedestrian crosses the road from the right. **E** A vehicle crosses the median to make a U-turn. **F** An abandoned wheelbarrow in the middle of the road

longer than 8 s. For hazardous videos longer than 11 s, we extracted the no-hazard videos from the same source videos, using segments of the video that did not overlap with the 8-s hazardous videos. These segments did not contain hazards, providing hazard and no-hazard videos that were matched in conditions as much as possible. Taking hazard and no-hazard videos from the same source videos allowed us to control for factors such as location, time of day, weather, road condition, vehicle type, and dashcam setup across the two video categories. This procedure produced 434 hazard and 316 no-hazard video clips.

Because the contents of the videos can change moment to moment, we expected the hazardousness of road situations to change quickly around hazard onsets. Therefore, to obtain the most accurate hazardousness ratings, we trimmed the 750 source videos into brief clips for participants to rate, resulting in 1356 clips for the rating experiment (see Stimuli).

Annotation procedure

Videos were initially annotated by a driver with at least 2 years of driving experience; then it was validated by another driver with over 15 years of driving experience. Points of disagreement were discussed until the raters reached consensus.

Temporal annotations

For each video, we annotated the frame on which the first visible deviation from normal state can be seen. We defined this as when the hazardous object first deviates from a non-threatening trajectory, such as a car starting to veer into the driver's lane. Prior to this point, there is no visible indication that the object is dangerous. A hazard did not have to enter the driver's path of travel to be coded as the first visible deviation. Sometimes, the first point of deviation corresponds to the first time the hazard is visible, like when an animal runs towards the road. Other times, an object can be visible for a considerable period before any visible signs of danger are seen, such as a bumper suddenly falling off the back of a car. The bumper would only be hazardous when it starts to fall off, despite being visible long before the hazardous event.

Notably, our definition of a hazard does not include latent hazards (objects or locations in the road scene that are more likely to evolve into hazards than the rest of the scene; Crundall et al., 2012; Vlakveld et al., 2011) because latent hazards often do not develop into a collision or near collision, since whether the latent hazard requires a response depends on the driver's assessment and may vary among individuals, making it more difficult to design behavioral experiments on hazard detection surrounding them, which was our reason for developing this dataset. Furthermore, the drivers' ability to identify latent hazards vary with experience (Crundall et al., 2012) and it is unclear to what extent individual drivers can agree upon which objects and locations are identified as latent hazards in a scene in real on-road scenarios under realistic time constraints.

Given that our motivation for developing this dataset is to facilitate behavioral experiments about hazard detection and dynamic scene perception more generally, we have chosen to focus our annotations on immediate hazards that require a vehicle maneuver to avoid a collision, rather than latent hazards. Although reacting to latent hazards would increase the margin of error and reduce the likelihood of an eventual collision, whether a vehicle maneuver is required depends on the individual driver's subjective judgement of the likelihood of a hazard occurring. Furthermore, any object in the driver's environment can be a potential hazard, but a driver would not necessarily react to all of them as if they were immediate hazards. Doing so could even be maladaptive, producing erratic behavior that would be dangerous to other road users. For example, it is possible that a car traveling in an adjacent lane may change lanes without signaling, heading on a collision trajectory. However, the driver would likely not pay special attention to a particular car until it starts to deviate from its lane because most cars travelling in the adjacent lane remain in their lanes. Therefore, the underlying assumption to the current method is that, given that the hazards requiring immediate responses in the dataset are unexpected, drivers are unlikely to pay special attention to the object that would eventually be involved in the near collisions until the first sign of deviation from the norm, even if doing so would improve their performance and road safety more generally. For these reasons, the annotations in this dataset are focused on immediate hazards that unfolded into near-collisions, and would have been collisions if a driver did not respond. The distribution of hazard onset times based on this definition is shown in Fig. 2A.

We also annotated the time when the driver made their first response to the hazard. This is defined as the first point at which the driver's response is visible in the footage, such as when the driver slowed down (braking) or began to swerve to evade a hazard. These response annotations were based on visible response only; braking responses can be inferred by closely observing the footage and watching for the vehicle's hood to dip towards the roadway. Swerving responses can similarly be inferred by looking for shifts in camera viewpoint. The distribution of response timing can is shown in Fig. 2B and C. Consistent with previous research on surprising events on the road, most driver responses occurred within 1.5 s of hazard onset (Green, 2000).

Categorical annotations

Videos were also categorized based on other potentially relevant characteristics, such as time of day, whether the video

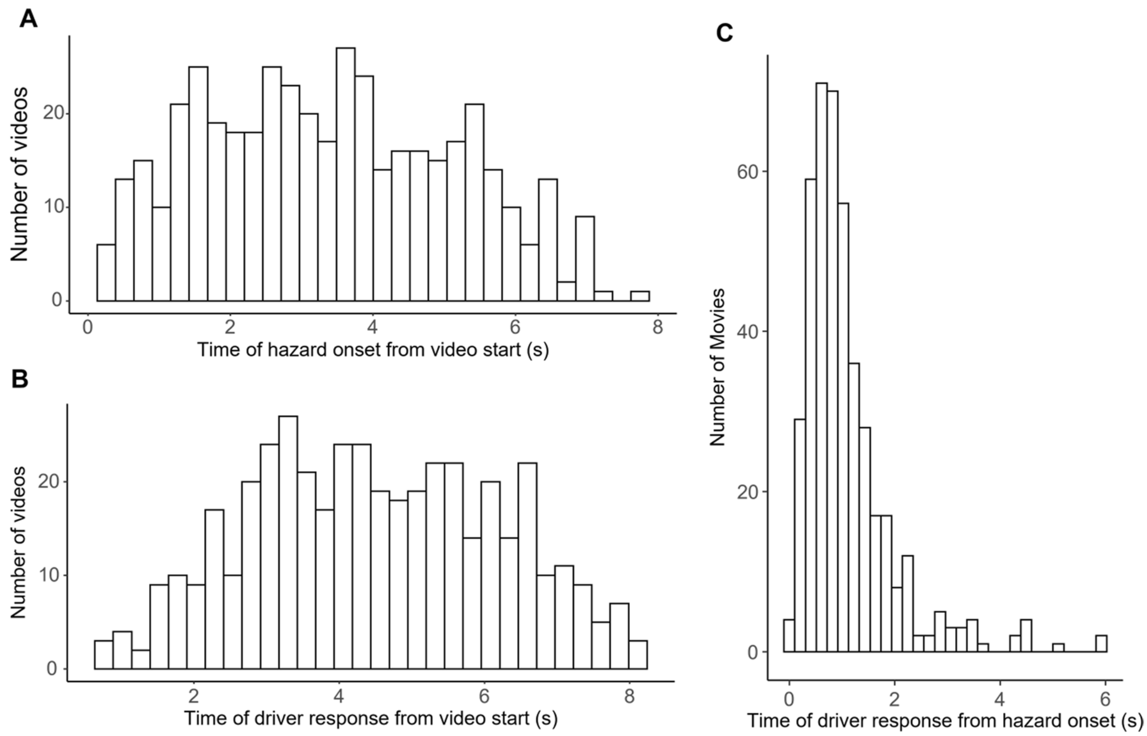

Fig. 2 **A** Histogram of hazard onset times relative to the beginning of the video. **B** Histogram showing the time of driver response from the beginning of each video. **C** Histogram of annotated driver response times relative to annotated hazard onset times. Most responses contained in the videos occur within 2 s of the annotated hazard onset

shows left- or right-hand traffic, and the type of hazard, such as pedestrian or vehicle hazard. Tables 1 and 2 show the distribution of videos for a selection of categorical annotations (see Appendix Tables 5, 6, 7 and 8 for a list of all annotations). Given that video footage was taken during naturalistic driving, hazards took many forms (see Fig. 1 for examples in each hazard category). This type of variability in the dataset was unconstrained to ensure that the dataset represents the natural variability in hazardous situations that drivers may need to respond to on the road. By categorizing these hazards, and the circumstances in which these hazards happen, we aim to make the dataset maximally useful for its users.

Spatial annotations

The locations of the hazards were also annotated in the duration between hazard onset and the driver's first visible response using rectangular bounding boxes. These annotations are useful for studies that require knowing the location of hazards, for example, for cueing studies (i.e., Wolfe et al., 2021). Secondly, this spatial information may be useful for eye-tracking studies as well, such as for area of interest analyses (Ahlström et al., 2021). The location of the dashboard, car, and other parts of the video that do not contain road information were annotated on the first frame of hazard onset. These annotations indicate areas of the video where no useful information for detecting hazards is present. Since the location of the dashcams relative to the vehicles do not change, these areas also stay the same for the duration of the video.

Participants

To obtain hazardousness ratings, we recruited 48 licensed drivers online from Prolific who were aged 20–35 (M = 29, SD = 4.35), with self-reported normal or corrected-to-normal vision. Half of the participants resided in North America, and the other half in the United Kingdom. Gender (cis and trans men and women) was balanced within each group. 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. Participants completed the study in two sessions. We compensated participants with 6.75 Great British Pounds (GBP; approximately 8 USD) per session. Participants also received a bonus of 8.50 GBP for completing both sessions, for a total of 22 GBP (approximately 27 USD) for completing the study.

Stimuli and materials

The experiment was programmed in PsychoPy/PsychoJS (Bridges et al., 2020; Peirce et al., 2019) and hosted online on Pavlovia. Participants completed the study on their own desktop or laptop computer, and the experiment was disabled on mobile and tablet devices.

For collecting rating data, we extracted 1356 brief 333-ms excerpts from the hazard and no-hazard videos using the procedure listed in the next paragraph. Videos were kept brief because driving situations can change within a fraction of a second, and we wanted to examine judgements of hazardousness at the scene gist level rather than average hazardousness across a longer video composed of multiple events. The 333-ms duration was chosen to mimic the time pressure of real driving because it was approximately 150% of the threshold duration that young-adult drivers needed to accurately detect a hazard, as determined in a previous experiment that used a subset of the same stimuli (Wolfe et al., 2020). All video clips were shown in between randomly generated noise masks. Each mask consisted of a grid of 36 × 64 squares, 20 pixels on each side, with a random grayscale intensity.

For the hazard-present videos, clips were taken from the 333 ms immediately preceding the annotated time of the driver's response (see Annotation procedure). Our rationale was that including any video segments after the driver's response would introduce visual cues that were not directly produced by the hazard itself (e.g., sudden changes in optic flow due to the vehicle swerving), and we were primarily interested in the detection of the hazard itself rather than the response. The 316 hazard-absent videos were trimmed into 333-ms segments that were, for most videos, extracted from at least 10 s before or after the annotated time of the hazard onset when possible. In addition, the hazard-absent clips were separated from one another by at least 3 s. Based on the available video segments, this process resulted in 432 hazard-present and 924 hazard-absent clips for collecting ratings data, for a total of 1356 clips.

Driving history questionnaire

We also administered a driving history questionnaire administered via an online survey service, Qualtrics, to measure driving experience. Although our sample included only licensed drivers, the population of licensed drivers captures a vast range of on-road experience, which we aimed to capture using a questionnaire. The driving history questionnaire is adapted from the driving history questionnaire used in the SHRP2 Naturalistic Driving Study (Campbell, 2012). The full questionnaire is available on OSF, and a summary of select statistics are available in Table 8 in the Appendix.

Procedure

On each trial, participants were shown the sequence of a noise mask for 250 ms, followed by the video clip for 333 ms, and then a second noise mask of 250 ms. Participants were then asked to indicate the hazardousness of the clip on a continuous slider from 0 (very safe) to 1 (very dangerous). Each observer saw and rated every single movie clip once in a randomized order. The rating response was untimed, and participants could rate videos at their own pace. There were 1356 trials in total (one per clip, consisting of 432 hazard and 924 no-hazard clips), which participants completed across two 45-min sessions on separate days. Participants were given the opportunity to take a break every 60 trials. At the end of the second session, participants completed a demographics and driving history questionnaire.

In addition, as a data quality measure, we included 40 catch trials in each session (approximately 6% of all trials) that were interleaved with the road hazard videos. On the catch trials, participants viewed a 1-s video clip of a non-driving scene and indicated whether the video was of an indoor or outdoor scene by pressing one of two keys on their keyboard. These videos were sourced from YouTube and public-domain stock footage websites and depicted a range of everyday settings and activities (e.g., cleaning, hiking, gardening, cooking). The clips were selected to be unambiguously indoor or outdoor. All participants met our inclusion criteria based on performance on the catch trials, achieving a minimum accuracy of 85%. The mean accuracy across all participants was 98.3%.

Data analysis

All analyses were conducted in R version 4.1.3. The main movie-wise analyses and the correlations conducted on driving-experience survey data and ratings data were pre-registered (see https://osf.io/52zes). We diverged from the pre-registration to conduct exploratory analyses on the effect of demographic variables on hazardousness ratings. We used zero- and one-inflated beta (ZOIB) regression modeling rather than the aligned rank transform ANOVA that we had pre-registered because there were many ratings that were exactly 0 or 1. This would have made the aligned rank transform ANOVA inappropriate due to the large number of ties after rank transformation (Luepsen, 2017). All ZOIB regression modeling was done using the *brms* package version 2.18.0 for R.

Movie-wise analysis

For movie-wise analyses, we report median hazardousness ratings to minimize the influence of outliers, and because the ratings were not normally distributed across participants due to the bounds at 0 and 1. To examine the extent to which observers agreed on the hazardousness of movies, we analyzed intra-class correlation (ICC) among raters. We used a mean-rating (k = 48), consistency, two-way random-effects model. We picked this ICC model because each participant may have a different definition of the midpoint of the scale. We were interested in how participants rated videos relative to other videos, rather than the absolute ratings, because absolute ratings may be biased by individual differences. This analysis was conducted using the *psych* package for R, version 2.2.9 (Revelle, 2023).

Participant-wise analysis

We examined several demographic variables to examine whether they could explain any individual differences in hazardousness ratings. For categorical demographics variables, we used ZOIB regression modeling for significance testing because we had a non-negligible amount of hazardousness ratings of zero and one, which could not be modeled by a beta regression model alone. Zero- and one-inflation allows us to model the probability of observing ratings of zeros and ones.

We examined gender because it is a factor that historically have been an interest to researchers (Eustace & Wei, 2010; Li et al., 1998; Massie et al., 1995; Tavris et al., 2001). The ZOIB model included gender and video type as fixed predictors and participant as a random predictor. The dependent variables were the ratings responses, and 0 and 1 inflation rates. These model estimates were then used to compute contrasts to examine whether gender differences exist for the rating of hazardous and no-hazard videos, as well as whether gender modulates the difference between hazard and non-hazard videos.

We also examined whether country of residence and left or right-hand driving interacted to affect hazardousness ratings. This is because we expected familiarity with a traffic system may allow participants to identify hazards, and therefore affect the hazardousness ratings. For example, a UK resident who is familiar with driving on the left side of the road may rate videos showing left-side driving differently than a North American (NA) resident who have never driven on a left-side driving road. To examine the effect of country of residence on hazardousness ratings, we fitted a zero- one-inflated beta regression model to the data. Individual responses were the dependent variable, and video type (hazard or no hazard), country of residence (NA vs. UK), road convention (left- or right-hand traffic), and all possible interactions were included as predictors. Participant was also included as a random predictor. The same model was used to model the zero- one-inflation rate and the conditional one inflation rate.

Because age and experience are closely linked, we also examined driving experience measures and how they correlated with hazardousness ratings. To measure driving experience, we adapted the driving history questionnaire from the SHRP2 naturalistic driving study (Campbell, 2012). Then, we correlated the quantitative items from the questionnaire with movie ratings to examine whether driving experience and history influenced hazardousness ratings. We used the Bonferroni correction for multiple comparisons, such that critical alpha is 0.005, (0.05 family wise alpha divided by 11 tests).

Results

Sample

We recruited 53 participants who attempted the experiment on Prolific. Three participants failed to complete session one, and two participants did not return for session two, resulting in 48 participants in our sample.

Movie-level analyses

Median ratings for each movie are distributed across the entire scale, consistent with the idea that drivers assess hazards in a more granular fashion than can be captured with a simple binary hazard present/absent designation. Figure 3 shows examples of videos from the highest and lowest quartiles of median hazardousness for each video type. The distribution of the ratings is bimodal, with no-hazard videos rated lower (mean rating = 0.193) and hazard videos rated higher (mean rating = 0.638). A Wilcoxon ranked test indicated that there was a significant effect of video type on hazardousness ratings ($V = 4656$, $p < 0.001$). The mean rating (k = 48), consistency, two-way random-effects ICC model showed high interrater agreement with an estimate of 0.987 (95% CI = [0.986, 0.988], $F(1355,63685) = 77.3$, $p < 0.001$).

The standard deviation of the ratings is shown as a function of the median rating for each movie in Fig. 4. The data form an inverted "U" shape: as expected, given the bounded nature of the ratings, the standard deviation at the ends of the scale is smaller than the middle by virtue of only having a single direction to vary. These data also suggest that the ratings near the middle are more ambiguous than the movies that are rated on the more extreme ends of the scale, as there tended to be more variation in movie ratings across observers for movies with moderate ratings.

Analyses of demographics

Effect of gender

Contrasts performed to examine the effect of gender on median ratings found small and negligible effects (see Table 3 for statistics). The interaction between Rater Gender and Video Type was the only contrast with a 95% confidence interval that did not cross zero, but the effect is so small (– 0.02) as to be negligible.

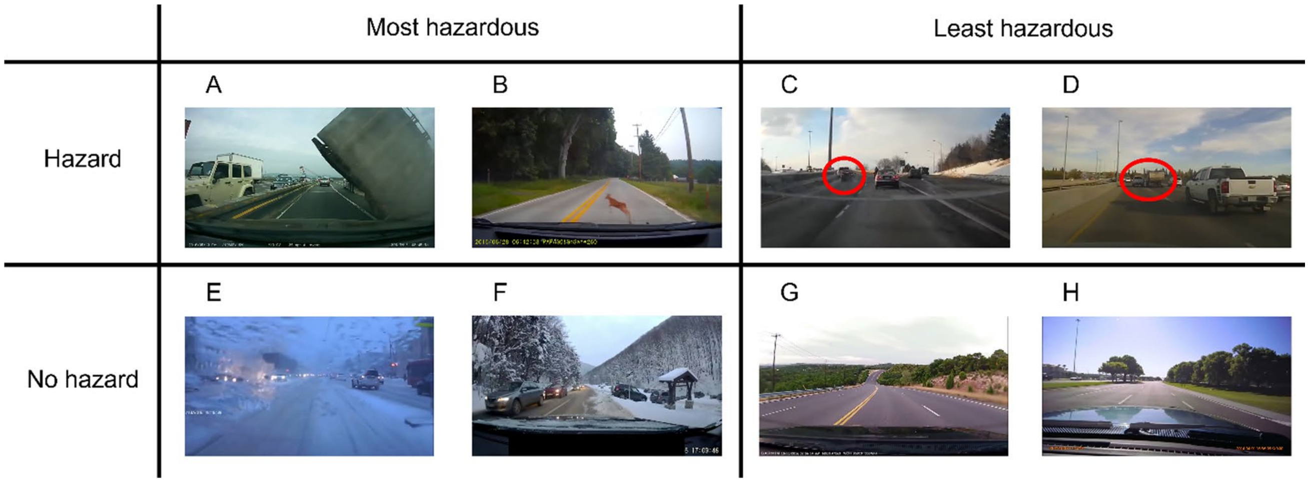

Fig. 3 Examples frames from videos that were rated as the most hazardous quartile (*left column*) and the least hazardous quartile (*right column*) for hazard-present (*top row*) and hazard-absent (*bottom row*) videos. Hazard-present videos rated as highly hazardous are typically videos of near collisions where the driving speed is usually fast, and the hazard is clearly visible in the footage (**A** A truck rolls over into the camera car's lane; **B** A deer runs across the road). The hazard-present videos rated as the least hazardous usually contained hazards that are more visually subtle than in the highly hazardous videos, usually due to the fact that hazards were more distant or smaller in size (**C** Two cars to the left of the camera car's lane collide with each other; **D** Two cars collide into each other in front of the camera car. Images C and D were enlarged and cropped, and hazards are *circled in red* for illustrative purposes. In the actual stimulus, the hazards were smaller and more of the road scene was visible, and there were no red circles. All other images in Fig. 4 are unaltered). For hazard-absent videos, the most hazardous videos usually contained unsafe driving conditions without requiring an immediate vehicle maneuvering response, such as adverse weather conditions (**E**), or when there are many potential, latent hazards (**F**). The no-hazard videos rated as least hazardous usually show clear visibility and an empty road (**G**, **H**)

Effect of country of residence

Figure 5 shows the ratings separated by country of residence, video type, and road convention. The results of the ZOIB regression showed that aside from the effect of video type, all other effects were negligible (see Table 1 for the full list of statistics). There was a main effect of video type: on average, no-hazard videos were rated as 0.37 arbitrary units less hazardous than videos containing hazards (95% CI = [– 0.38, – 0.37]). Overall, right-hand traffic videos were rated as slightly less hazardous than left-hand traffic videos (estimated effect = – 0.01, 95% CI = [– 0.02, – 0.02]), and there was a negligible main effect of country of residence (estimated effect = 0, 95% CI = [– 0.01, 0.01]). Country of residence had a very small modulatory effect on video type – UK residents rated safe and hazardous videos as slightly more similar than NA residents (estimated effect = – 0.05, 95% CI = [– 0.06, – 0.04]). Additionally, the interaction between country of residence and road convention was negligible (estimated effect = 0, 05% CI = [– 0.01, 0.01]). There was no three-way interaction among country of residence, video type, and road convention. The full model coefficients are reported in Table 6 in the Appendix. In summary, there was a large effect of video type on hazardousness ratings, and the differences between left- and right-hand traffic and between North America and UK residents were negligible.

Driving history

Next, we examined the relation between driving experience and the ability to distinguish hazard from no-hazard videos from hazardousness ratings (see Table 8 in the Appendix for a list of quantitative experience-related variables). To measure a driver' ability to distinguish hazard videos and non-hazard videos, we calculated the difference between the average rating for hazards videos and the average rating for non-hazard videos for each participant. Then, we calculated Spearman's correlation coefficient between questionnaire items and the rating difference. None of the correlations were statistically significant after correcting for multiple comparisons (see Table 4).

Discussion

The Road Hazard Stimuli is a rich video database (available at https://osf.io/tgzb7) of naturally occurring road situations which have been extensively annotated to make them useful for experimental work in driver behavior, visual perception, and other applications. Hazardous situations on the road are, fortunately, very rare and can take a large range of forms. The typical collision rate for drivers is estimated to be once every 10 years (Horswill et al., 2021). Based on this

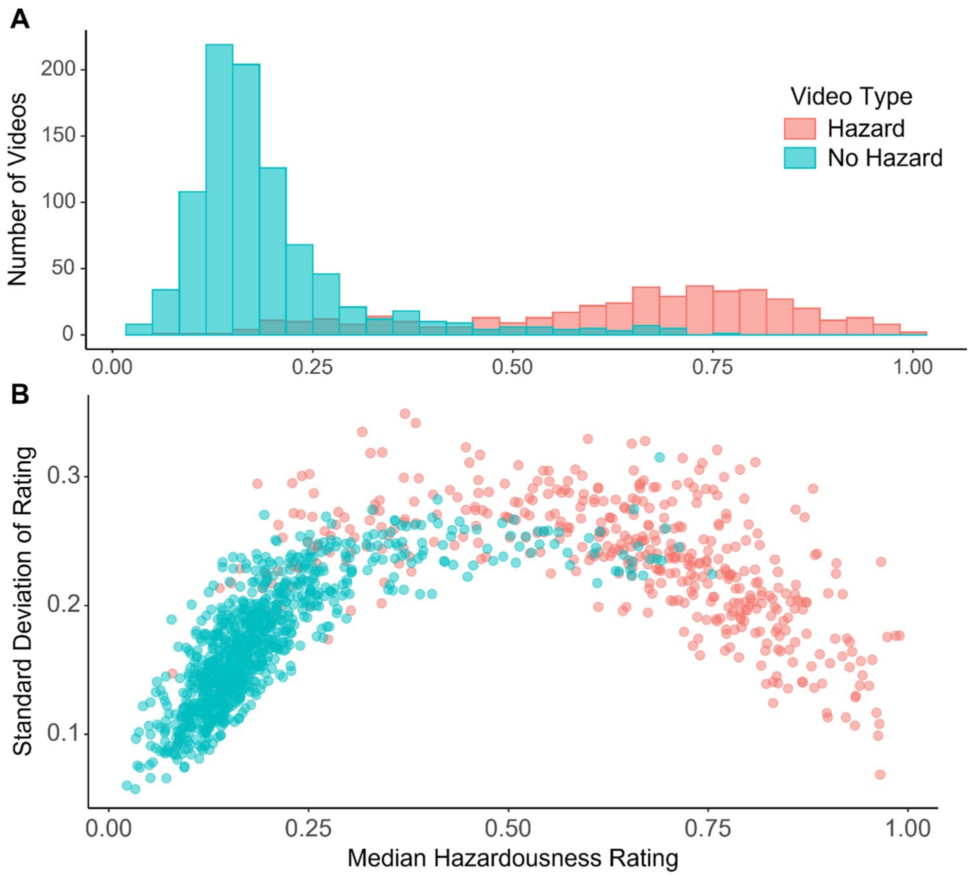

Fig. 4 **A** Histogram of the distribution of median hazardousness ratings per movie. Turquoise bars represent movies without hazards, and coral bars represent movies with hazards. A rating of 0 represents very safe, and rating of 1 represents very dangerous. **B** Standard deviation of ratings plotted as a function of median rating for each movie. There was more variation among raters for movies rated towards the middle of the scale than movies with extreme hazardousness rating

Table 3 Results of contrasts of interest using estimates of zero- and one- inflated beta regression model. The dependent variable of all models were hazardousness ratings. To estimate the effect of gender (rows 1–3), the predictors included video type (hazard or no hazard) and gender (man or woman), and participants as a random effect. For all other rows, the ZOIB model's predictors included video type (hazard or no hazard), road convention (left- or right-handed traffic), and country of residence (NA or UK), with participants as a random effect. Text was bolded to highlight contrasts for which the 95% confidence interval does not cross zero

Contrast	Estimate	95% CI
Gender effect (man - woman) within hazard videos	– 0.03	[– 0.09, 0.03]
Gender effect within no-hazard videos	– 0.01	[– 0.06, 0.03]
Gender effect on video type effect (hazard – no hazard)	**– 0.02**	**[– 0.03, 0.00]**
Main effect of country of residence	0.02	[– 0.3, 0.06]
Main effect of video type	**– 0.37**	**[– 0.38, – 0.37]**
Main effect of road convention	**– 0.01**	**[– 0.02, – 0.01]**
Country of residence × video type	**– 0.05**	**[– 0.06, – 0.04]**
Country of residence × road convention	0.00	[– 0.01, 0.01]
Country of residence × video type × driving	0.01	[– 0.01, 0.03]

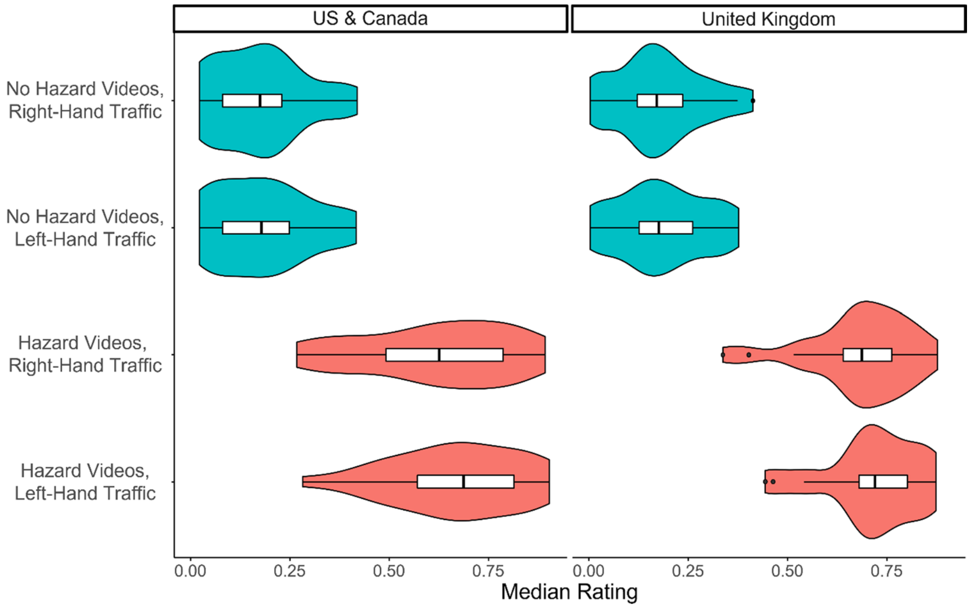

Fig. 5 Violin plot of median hazardousness rating for each movie categorized according to whether the video contains a hazard with inset boxplots. *Turquoise* represents videos that do not contain hazards and *red* represents videos that contain hazards. Raters were divided into residents of North America (*left half*) and residents of the United Kingdom (*right half*) before median ratings were calculated for each video. The *bolded vertical line* indicates the median, and the *white rectangle* represents the middle 50% of the data. Outliers are represented by *dots* which lie outside 1.5 times the interquartile range

estimate, a driver would only encounter a miniscule fraction of the situations incorporated in the dataset over their lifetime of driving. The stimulus set also includes extensive annotations, allowing researchers to take what they need, whether or not they are asking driving-related questions.

We found that despite the inherent variability of hazards, our drivers were remarkably consistent in their assessments of what was or was not dangerous to them. Moreover, driving history and driver demographics had very weak effects on drivers' assessments. Due to the constraints of online studies, we were unable to precisely control for the size of the videos shown to participants. This may have made the data we collected noisier than would be if done in a controlled laboratory setting, such as the ability to detect ambiguous hazards such as in Fig. 3C, D. However, given the high inter-rater agreement of the ratings, and the small magnitude of effects of driving history and driver demographics, data collected in a more controlled laboratory setting is unlikely to change our findings. These findings suggest that studies built around our stimuli can ask a vast array of questions and can speak to driver behavior more broadly. This dataset provides an agreed-upon benchmark, and lays a foundation for a wide range of future research in driver behavior, visual perception and beyond.

Table 4 Results of Spearman correlation analyses for each questionnaire response and the mean difference in rating between hazard and non-hazard videos. For each test, the degree of freedom was 43, and the critical p value was 0.005 after correcting for multiple comparisons. In general, the correlations are small to moderate and none of them survived the correction for multiple comparisons

Measure	Spearman's ρ	p value
Year started	0.04	0.79
Current driving frequency	0.05	0.73
Mileage last year	0.15	0.34
Years owned vehicle	0.26	0.09
Age of licensure	– 0.10	0.50
Last eye exam	0.13	0.39
Auto insurance last year	– 0.03	0.86
Mileage in last 5 years	0.12	0.45
Traffic violations	0.89	0.56
Number of crashes	0.29	0.05
Age	– 0.04	0.70

Given that our sample consisted of young adults who are moderately experienced drivers, there are many questions that this stimulus set could help answer. One is the effect of driving training and experience on detecting dangerous situations. Although an untimed rating task is not sensitive to differences in driver experience (Crundall et al., 1999), novice drivers are poorer at responding to hazards when processing time is restricted (Jackson et al., 2009). Novice drivers also experience higher collision risk (Cooper et al., 1995; Gregersen & Bjurulf, 1996; Mayhew et al., 2003), and are less able to predict future hazards compared to expert drivers (Deery, 1999; Renge, 1998; Wallis & Horswill, 2007). Designing a task around these abilities, such as detecting latent hazards, or objects and locations that are likely to become dangerous, may be a more fruitful way to identify differences between novice and experience drivers. Given that some videos without immediate hazards were given ratings above zero, this dataset may also be useful for studying the detection of latent hazards. Understanding what novice drivers need to learn to become experienced drivers is a crucial step for developing more efficient training regimes aimed to reduce collision risk during this particularly vulnerable period that every driver must experience. This dataset can be used as a principled starting point for developing such training regimes aimed at improving hazard detection in novice drivers.

Additionally, previous studies have found regional differences in drivers' hazard detection abilities, presumably due to different hazard prevalence rates among countries (Di Stasi et al., 2020; Sivak et al., 1989; Ventsislavova et al., 2019). However, regional effects depends on the design of the study used and the populations sampled (Bazilinskyy et al., 2020; Ventsislavova et al., 2019). Although we observed relatively small effects of demographic variables on ratings of hazardousness, future research with a broader sample may yield larger effects.

What can this stimulus set be used for?

This dataset provides immense flexibility in the types of experiment designs. The standard format of the videos makes it easy to systematically manipulate many aspects of the stimuli. The ability to edit and transform the videos programmatically (e.g., trimming the videos to shorter segments, blurring or down sampling the videos) expands the types of questions that researchers can ask. Several studies have already used these videos to study hazard detection in human drivers. For example, the videos have been used to examine the shortest duration of video required to accurately detect hazards in young and older drivers by varying video duration across trials (Wolfe et al., 2020). The study found that to detect hazards with 80% accuracy, young drivers need only approximately 220 ms of preview, whereas older drivers require almost double the preview duration.

These stimuli also have been used to examine whether and how phenomena commonly found in laboratory settings apply to the perception of dynamic road scenes. For example, Kosovicheva et al. (2023) demonstrated the low prevalence effect when drivers were asked to detect overt hazards in road scenes – observers miss rare hazards more frequently than common hazards. However, these are far from the only driving-related questions that these stimuli can help answer; effects of reduced visual acuity (Guidi et al., 2022), the potential role of attentional cueing (B. Wolfe et al., 2021) and questions of how drivers use peripheral vision can also be investigated using these stimuli.

The annotations provide useful temporal, spatial, and categorial information about the videos which provide researchers with the ability to pick and choose the information that they need for the question they are studying. For example, spatial annotations can fast track developing eye-tracking studies, which are useful for understanding how observers sample dynamic scenes, and why drivers look where they do. To make sense of eye-tracking data, driver eye movements must be spatially mapped to what the driver is seeing (Ahlström et al., 2021). Spatial annotations provide the locations of hazards and informative regions of the video and will make it much easier to map where people looked with what they were looking at. For example, eye-tracking data can help identify reasons why drivers miss hazards, such as looking in the incorrect location, or looking at the correct location but failing to recognize a hazard (Wolfe et al., 2022). The annotations can facilitate examining these questions, which will deepen our understanding of how drivers understand their visual environment, and scene perception more broadly.

Although this stimulus set contains only visual information, it can be a starting point for investigating how other sensory domains contribute to road scene perception. Information from other modalities can be presented along with the videos in tandem. Given the temporal annotations, they can be yoked to hazard onset, or driver response. For example, we are currently investigating the effect of auditory cue timing and accuracy on road hazard detection, with the goal to determine whether the benefit of accurate auditory cue outweighs the detrimental effects of potential invalid auditory cue. These stimuli may be useful for studying multi-sensory integration in an on-road context, which may inform the design of in-car technology and driver assistance systems.

Finally, although this dataset was created for behavioral research in mind, it contains rich data that may be useful for other types of work. For example, driver responses were also captured and annotated in this video set, which may be useful for computer vision models aimed at predicting driver behavior based on the driver's visible environment. The annotations by humans provided in the dataset can be used to validate machine learning performance. By making this dataset publicly available, we hope that researchers in other fields can also use it for purposes beyond our original intentions.

As we note earlier, our goal in developing this stimulus set was to capture the range of variability of hazards that drivers may encounter on the road. However, we also acknowledge that there are a number of trade-offs in taking this approach. For instance, videos posted to social media may not be fully representative of hazards that drivers encounter on the road. The trade-off between variability and representativeness is one that the experimenter must carefully balance depending on the nature of their study. We also note that the hazardousness ratings for the hazard-present videos span a wide range on the rating scale, so while some of the videos posted to social media may be rated towards the high end of the scale, many of the videos were also rated lower. By providing the ratings for each video, researchers may select videos that are more or less extreme, depending on the goals of their particular study. In addition, the videos we selected were taken without regard to country of recording, often because that information is not directly available from videos posted on social media. For studies that aim to draw conclusions about specific road environments, experimenters may wish to develop more focused stimulus sets; making such stimulus sets widely available would also benefit those who study particular road environments.

This video set contains only front-facing footage, which may not fully capture the demands of real-world driving because drivers are also required to occasionally monitor the side, rear-view mirrors, and the vehicle's dashboard console on the road. Furthermore, more peripheral information about the environment contributes to the detection of hazards (Shahar et al., 2010), which is not captured in this dataset. Researchers interested in how drivers divided attention among multiple displays can still use this dataset while simulating other displays.

Conclusion

The videos presented here represent a collection of road videos in a standardized format. The categorical, temporal, and spatial annotations and the hazardousness ratings provide rich descriptive information about the dataset, and can be used to ask a wide array of questions from applied hazard perception and driving safety topics to fundamental topics such as the guidance of eye movements in dynamic scenes, and computer vision. Although this dataset was designed for studying the perception of road-scenes, it is also a useful tool for behavioral research and beyond.

Appendix

Table 5 List of all annotations in the dataset

Annotation name	Description
MovieFile	Name of the file
SourceVid	Name of the source video
left_or_right	Vehicles drive on the left- or right-hand side of the road
VideoDuration	Duration of the video
Hazard	Whether a hazard is present in the video
day_or_night	Time of day the video is taken
Temporal annotations	
Main Response	First driver response observed in video
Driver_Response	Time in seconds of first driver response from start of video
cue_time_FINAL	Time in seconds of hazard onset from the start of video, i.e., first sign that the situation is hazardous, not when the object that will eventually be hazardous is first visible
Cue-Response Delta	Time in seconds between hazard onset and driver response
event_type	Detailed hazard type
target_type	Category of hazard: vehicle, pedestrian, cyclist, animal
hazardSide	The location of the hazard relative to video midline
Spatial annotations	
Video width	Pixel width of video
Video height	Pixel height of video
Video fps	Frame rate
Frame number	Frame number from beginning of video
Frame time (s)	Time from beginning of video
Annotation type	Object being annotated
Point #	Point number
X-coordinate	X position of the point in pixels, 0 at left edge
Y-coordinate	Y position of the point in pixels, 0 at bottom edge

Table 6 All estimated coefficients of gender model

Model details:

Family: zero_one_inflated_beta

Links: mu = logit; phi = log; zoi = logit; coi = logit

Formula: Hazardousness rating ~ Video type * Gender + (1 | Participant)

phi ~ 1

zoi ~ Video type * Gender + (1 | Participant)

coi ~ Video type * Gender + (1 | Participant)

Number of observations: 65088

Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;

total post-warmup draws = 4000

	Estimate	Est. Error	95% CI	Rhat	Bulk ESS	Tail ESS
Intercept	0.4	0.09	[0.22, 0.58]	1.01	320	666
Phi Intercept	1.17	0.01	[1.16, 1.18]	1	8027	3160
Zoi Intercept	– 7.54	0.76	[– 9.17, – 6.17]	1	1334	1903
Coi Intercept	4.06	1.57	[1.29, 7.62]	1	2968	3064
Video Type No Hazard	– 1.53	0.01	[– 1.55, – 1.5]	1	6059	3382
Gender Woman	0.12	0.13	[– 0.12, 0.37]	1.02	282	806
Video Type Safe × Gender Woman	– 0.06	0.02	[– 0.1, – 0.03]	1	6440	3210
Zoi Video Type No Hazard	– 0.17	0.13	[– 0.43, 0.09]	1	7310	2721
Zoi Gender Woman	– 1.02	1.06	[– 3.06, 1.12]	1	1170	1991
Zoi Video Type No Hazard × Gender Woman	0.21	0.34	[– 0.44, 0.86]	1	6714	3195
Coi Video Type No Hazard	– 9.71	1.54	[– 13.18, – 7.22]	1	4556	2650
Coi Gender Woman	– 0.79	2.62	[– 5.79, 4.68]	1	1892	1999
Coi Video type Safe × Gender Woman	1.56	2.95	[– 5.03, 6.64]	1	2390	1867

Table 7 All estimated coefficients of country of residence regression model. Model details:

Family: zero_one_inflated_beta

Links: mu = logit; phi = log; zoi = logit; coi = logit

Formula: slider.response ~ Video type * Country of residence * Road convention + (1 | Participant)

phi ~ 1

zoi ~ Video type * Country of residence * Road convention + (1 | Participant)

coi ~ Video type * Country of residence * Road convention + (1 | Participant)

Number of observations: 65088

Draws: 4 chains, each with iter = 4000; warmup = 1000; thin = 1;

total post-warmup draws = 12000

	Estimate	Est. Error	L 95% CI	U 95% CI	Rhat	Bulk ESS	Tail ESS
Intercept	0.51	0.08	0.33	0.63	1.24	12	22
Phi Intercept	1.18	0.01	1.17	1.19	1.02	207	457
ZOI Intercept	– 7.74	0.85	– 9.35	– 6.16	1.05	48	124
COI Intercept	11332.5	4511.21	3170.17	20106.41	2.22	5	13
Video type Safe	– 1.56	0.03	– 1.61	– 1.5	1.04	87	255
Country of residence UK	0.18	0.11	– 0.05	0.38	1.46	8	31
Road convention Right-hand traffic	– 0.13	0.02	– 0.17	– 0.08	1.06	79	172
Video type Safe × Country of residence UK	– 0.21	0.04	– 0.28	– 0.14	1.07	74	255
Video type Safe × Road convention Right-hand traffic	0.14	0.03	0.08	0.19	1.05	78	229
Country of residence UK × Road convention Right-hand traffic	0.01	0.03	– 0.05	0.08	1.09	49	258
Video type Safe × Country of residence UK × Road convention Right-hand traffic	– 0.04	0.04	– 0.11	0.04	1.08	64	278
ZOI Video type Safe	– 0.89	0.41	– 1.64	– 0.07	1.05	68	103
ZOI Country of residence UK	– 1.22	1.38	– 4.22	1.34	1.1	36	57
ZOI Road convention Right-hand traffic	– 0.07	0.31	– 0.65	0.57	1.04	80	135
ZOI Video type Safe × Country of residence UK	2.23	0.82	0.73	3.85	1.09	49	83
ZOI Video type Safe × Road convention Right-hand traffic	0.41	0.44	– 0.47	1.22	1.05	65	109
ZOI Country of residence UK × Road convention Right-hand traffic	0.58	0.75	– 0.75	2.06	1.07	59	90
ZOI Video type Safe × Country of residence UK × Road convention Right-hand traffic	– 1.19	0.86	– 2.89	0.4	1.1	46	91
COI Video type Safe	– 13274	5287.79	– 23348.9	– 3955.36	2.23	5	14
COI Country of residence UK	– 11333.7	4511.48	– 20109	– 3170.58	2.22	5	13
COI Road convention Right-hand traffic	– 11328.7	4511.39	– 20102.7	– 3165.04	2.22	5	13
COI Video type Safe × Country of residence UK	10772.68	4564.21	3178.52	20817.87	2.02	5	19
COI Video type Safe: Road convention Right-hand traffic	13264.8	5287.79	3945.71	23340.36	2.23	5	14
COI Country of residence UK × Road convention Right-hand traffic	11333.58	4511.6	3164.45	20110.9	2.22	5	13
COI Video type Safe × Country of residence UK × Road convention Right-hand traffic	– 11122.2	4697.59	– 21280.7	– 3235.56	2.02	5	19

Table 8 Summary of select quantitative results of driving history questionnaire. The full questionnaire and responses are available on OSF (https://osf.io/tgzb7)

Item	Median	Min	Max
Years since started driving	9	1	19
Licensure age (years)	18	15	30
Number of crashes in the past 10 years	0	0	1

References

AAA Foundation for Traffic Safety. (2021). *New American Driving Survey: Updated Methodology and Results from July 2019 to June 2020* (pp. 1–31). AAA Foundation for Traffic Safety https://aaafoundation.org/wp-content/uploads/2021/04/New-American-Driving-Survey-Report-April-2021-1.pdf. Accessed 1 Oct 2023

Ahlström, C., Kircher, K., Nyström, M., & Wolfe, B. (2021). Eye tracking in driver attention research—How gaze data interpretations influence what we learn. *Frontiers in Neuroergonomics, 2*, 778043. https://doi.org/10.3389/fnrgo.2021.778043

Bazilinskyy, P., Eisma, Y. B., Dodou, D., & de Winter, J. C. F. (2020). Risk perception: A study using dashcam videos and participants from different world regions. *Traffic Injury Prevention, 21*(6), 347–353. https://doi.org/10.1080/15389588.2020.1762871

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. https://doi.org/10.3758/s13414-013-0603-1

Borowsky, A., Shinar, D., & Parmet, Y. (2008). The relation between driving experience and recognition of road signs relative to their locations. *Human Factors*.https://doi.org/10.1518/001872008X288330

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*, e9414. https://doi.org/10.7717/peerj.9414

Campbell, K. L. (2012). *SAFETY The SHRP 2 Naturalistic Driving Study*. 8.

Chan, F.-H., Chen, Y.-T., Xiang, Y., & Sun, M. (2017). Anticipating Accidents in Dashcam Videos. In S.-H. Lai, V. Lepetit, K. Nishino, & Y. Sato (Eds.), *Computer Vision – ACCV 2016* (pp. 136–153). Springer International Publishing. https://doi.org/10.1007/978-3-319-54190-7_9

Cooper, P. J., Pinili, M., & Chen, W. (1995). An examination of the crash involvement rates of novice drivers aged 16 to 55. *Accident Analysis and Prevention, 27*(1), 89–104. https://doi.org/10.1016/0001-4575(94)00052-N

Crundall, D., Chapman, P., Trawley, S., Collins, L., van Loon, E., Andrews, B., & Underwood, G. (2012). Some hazards are more attractive than others: Drivers of varying experience respond differently to different types of hazard. *Accident Analysis & Prevention, 45*, 600–609. https://doi.org/10.1016/j.aap.2011.09.049

Crundall, D., Underwood, G., & Chapman, P. (1999). Driving experience and the functional field of view. *Perception, 28*(9), 1075–1087. https://doi.org/10.1068/p281075

Deery, H. A. (1999). Hazard and risk perception among young novice drivers. *Journal of Safety Research, 30*(4), 225–236. https://doi.org/10.1016/S0022-4375(99)00018-3

Di Stasi, L. L., Diaz-Piedra, C., Morales, J. M., Kurapov, A., Tagliabue, M., Bjärtå, A., ... & Catena, A. (2020). A cross-cultural comparison of visual search strategies and response times in road hazard perception testing. *Accident Analysis & Prevention, 148*, 105785. https://doi.org/10.1016/j.aap.2020.105785

Duivenvoorden, K., Hogema, J., Hagenzieker, M., & Wegman, F. (2015). The effects of cyclists present at rural intersections on speed behavior and workload of car drivers: A driving simulator study. *Traffic Injury Prevention, 16*(3), 254–259. https://doi.org/10.1080/15389588.2014.937484

Eustace, D., & Wei, H. (2010). The role of driver age and gender in motor vehicle fatal crashes. *Journal of Transportation Safety & Security, 2*(1), 28–44. https://doi.org/10.1080/19439961003590811

Falkmer, T., & Gregersen, N. P. (2001). Fixation patterns of learner drivers with and without cerebral palsy (CP) when driving in real traffic environments. *Transportation Research Part F: Traffic Psychology and Behaviour, 4*(3), 171–185. https://doi.org/10.1016/S1369-8478(01)00021-3

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research, 32*(11), 1231–1237. https://doi.org/10.1177/0278364913491297

Green, M. (2000). "How long does it take to stop?" Methodological analysis of driver perception-brake times. *Transportation Human Factors, 2*(3), 195–216. https://doi.org/10.1207/STHF0203_1

Gregersen, N. P., & Bjurulf, P. (1996). Young novice drivers: Towards a model of their accident involvement. *Accident Analysis and Prevention, 28*(2), 229–241. https://doi.org/10.1016/0001-4575(95)00063-1

Guidi, S., Ghuman, C., Kosovicheva, A., & Wolfe, B. (2022). Effects of Blur on duration thresholds for road hazard detection. *Journal of Vision, 22*(14), 4058. https://doi.org/10.1167/jov.22.14.4058

Hankey, J. M., Perez, M. A., & McClafferty, J. A. (2016). *Description of the SHRP 2 Naturalistic Database and the Crash, Near-Crash, and Baseline Data Sets [Technical Report]*. Virginia Tech Transportation Institute https://vtechworks.lib.vt.edu/handle/10919/70850

Horswill, M. S., Hill, A., Silapurem, L., & Watson, M. O. (2021). A thousand years of crash experience in three hours: An online hazard perception training course for drivers. *Accident Analysis & Prevention, 152*, 105969. https://doi.org/10.1016/j.aap.2020.105969

Jackson, L., Chapman, P., & Crundall, D. (2009). What happens next? Predicting other road users' behaviour as a function of driving experience and processing time. *Ergonomics, 52*(2), 154–164. https://doi.org/10.1080/00140130802030714

Kim, T., & Zhang, H. M. (2011). Interrelations of reaction time, driver sensitivity, and time headway in congested traffic. *Transportation Research Record, 2249*(1), 52–61. https://doi.org/10.3141/2249-08

Kosovicheva, A., Wolfe, J. M., & Wolfe, B. (2023). Taking prevalence effects on the road: Rare hazards are often missed. *Psychonomic Bulletin & Review, 30*(1), 212–223. https://doi.org/10.3758/s13423-022-02159-0

Langford, J., Charlton, J. L., Koppel, S., Myers, A., Tuokko, H., Marshall, S., ... & Macdonald, W. (2013). Findings from the Candrive/Ozcandrive study: Low mileage older drivers, crash risk and reduced fitness to drive. *Accident Analysis & Prevention, 61*, 304–310. https://doi.org/10.1016/j.aap.2013.02.006

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, 99*(14), 9596–9601. https://doi.org/10.1073/pnas.092277599

Li, G., Baker, S. P., Langlois, J. A., & Kelen, G. D. (1998). Are female drivers safer? An application of the decomposition method. *Epidemiology, 9*(4), 379–384.

Luepsen, H. (2017). The aligned rank transform and discrete variables: A warning. *Communications in Statistics - Simulation and Computation, 46*(9), 6923–6936. https://doi.org/10.1080/03610918.2016.1217014

Marshall, S. C., Man-Son-Hing, M., Bédard, M., Charlton, J., Gagnon, S., Gélinas, I., Koppel, S., Korner-Bitensky, N., Langford, J., Mazer, B., Myers, A., Naglie, G., Polgar, J., Porter, M. M., Rapoport, M., Tuokko, H., Vrkljan, B., & Woolnough, A. (2013a). Protocol for Candrive II/Ozcandrive, a multicentre prospective older driver cohort study. *Accident Analysis & Prevention, 61*, 245–252. https://doi.org/10.1016/j.aap.2013.02.009. Accessed 20 Feb 2023

Marshall, S. C., Wilson, K. G., Man-Son-Hing, M., Stiell, I., Smith, A., Weegar, K., Kadulina, Y., & Molnar, F. J. (2013b). The Canadian Safe Driving Study—Phase I pilot: Examining potential logistical barriers to the full cohort study. *Accident Analysis & Prevention, 61*, 236–244. https://doi.org/10.1016/j.aap.2013.04.002. Accessed 20 Feb 2023

Massie, D. L., Campbell, K. L., & Williams, A. F. (1995). Traffic Accident involvement rates by driver age and gender. *Accident Analysis and Prevention, 27*(1), 73–87. https://doi.org/10.1016/0001-4575(94)00050-V

Mayhew, D. R., Simpson, H. M., & Pak, A. (2003). Changes in collision rates among novice drivers during the first months of driving. *Accident Analysis and Prevention, 35*(5), 683–691. https://doi.org/10.1016/S0001-4575(02)00047-7

Moharrer, M., Tang, X., & Luo, G. (2020). With motion perception, good visual acuity may not be necessary for driving hazard detection. *Translational Vision Science & Technology, 9*(13), 18. https://doi.org/10.1167/tvst.9.13.18

Mourant, R. R., & Rockwell, T. H. (1972). Strategies of visual search by novice and experienced drivers. *Human Factors, 14*(4), 325–335. https://doi.org/10.1177/001872087201400405

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*(3), 353–383. https://doi.org/10.1016/0010-0285(77)90012-3

Oliva, A. (2005). Gist of the scene. *Neurobiology of attention* (pp. 251–256). Elsevier.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Renge, K. (1998). Drivers' hazard and risk perception, confidence in safe driving, and choice of speed. *IATSS Research, 22*, 103–110.

Revelle, William. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. https://CRAN.R-project.org/package=psych. Accessed 15 June 2023

SAE International. (2018). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. SAE International. https://doi.org/10.4271/J3016_201806

Saunier, N., Ardö, H., Jodoin, J.-P., Laureshyn, A., Nilsson, M., Svensson, Å., Miranda-Moreno, L., Bilodeau, G.- A., & Åström, K. (2014). *A public video dataset for road transportation applications: 93th TRB annual meeting*. Transportation Research Board 93rd Annual Meeting, Washington DC. https://portal.research.lu.se/en/publications/a-public-video-dataset-for-road-transportation-applications

Schall, M. C., Rusch, M. L., Lee, J. D., Dawson, J. D., Thomas, G., Aksan, N., & Rizzo, M. (2013). Augmented Reality Cues and Elderly Driver Hazard Perception. *Human Factors, 55*(3), 643–658. https://doi.org/10.1177/0018720812462029

Schieber, F., Schlorholtz, B., & McCall, R. (2009). Visual Requirements of Vehicular Guidance. In C. Castro (Ed.), *Human factors of visual and cognitive performance in driving* (pp. 31–50). CRC Press.

Shahar, A., Alberti, C. F., Clarke, D., & Crundall, D. (2010). Hazard perception as a function of target location and the field of view. *Accident Analysis and Prevention, 42*(6), 1577–1584. https://doi.org/10.1016/j.aap.2010.03.016

Sivak, M. (1996). The information that drivers use: Is it indeed 90% visual? *Perception, 25*(9), 1081–1089. https://doi.org/10.1068/p251081

Sivak, M., Soler, J., Tränkle, U., & Spagnhol, J. M. (1989). Cross-cultural differences in driver risk-perception. *Accident Analysis & Prevention, 21*(4), 355–362. https://doi.org/10.1016/0001-4575(89)90026-2

Tavris, D. R., Kuhn, E. M., & Layde, P. M. (2001). Age and gender patterns in motor vehicle crash injuries: Importance of type of crash and occupant role. *Accident Analysis & Prevention, 33*(2), 167–172. https://doi.org/10.1016/S0001-4575(00)00027-0

Theeuwes, J. (2021). Self-explaining roads: What does visual cognition tell us about designing safer roads? *Cognitive Research: Principles and Implications, 6*(1), 15. https://doi.org/10.1186/s41235-021-00281-6

Underwood, G., Crundall, D., & Chapman, P. (2002). Selective searching while driving: The role of experience in hazard detection and general surveillance. *Ergonomics, 45*(1), 1–12. https://doi.org/10.1080/00140130110110610

Ventsislavova, P., Crundall, D., Baguley, T., Castro, C., Gugliotta, A., Garcia-Fernandez, P., ... & Li, Q. (2019). A comparison of hazard perception and hazard prediction tests across China, Spain and the UK. *Accident Analysis & Prevention, 122*, 268–286. https://doi.org/10.1016/j.aap.2018.10.010

Vlakveld, W., Romoser, M. R. E., Mehranian, H., Diete, F., Pollatsek, A., & Fisher, D. L. (2011). Do crashes and near crashes in simulator-based training enhance novice drivers' visual search for latent hazards? *Transportation Research Record, 2265*, 153–160. https://doi.org/10.3141/2265-17

Wallis, T. S. A., & Horswill, M. S. (2007). Using fuzzy signal detection theory to determine why experienced and trained drivers respond faster than novices in a hazard perception test. *Accident Analysis & Prevention, 39*(6), 1177–1185. https://doi.org/10.1016/j.aap.2007.03.003

Wolfe, B., Fridman, L., Kosovicheva, A., Seppelt, B., Mehler, B., Reimer, B., & Rosenholtz, R. (2019). Predicting road scenes from brief views of driving video. *Journal of Vision, 19*(5), 8. https://doi.org/10.1167/19.5.8

Wolfe, B., Kosovicheva, A., Stent, S., & Rosenholtz, R. (2021). Effects of temporal and spatiotemporal cues on detection of dynamic road hazards. *Cognitive Research: Principles and Implications, 6*, 80. https://doi.org/10.1186/s41235-021-00348-4

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. https://doi.org/10.1037/xge0000665

Wolfe, J. M., Kosovicheva, A., & Wolfe, B. (2022). Normal blindness: When we look but fail to see. *Trends in Cognitive Sciences, 26*(9), 809–819. https://doi.org/10.1016/j.tics.2022.06.006

Yan, M. K., Kumar, H., Kerr, N., Medeiros, F. A., Sandhu, S. S., Crowston, J., & Kong, Y. X. G. (2019). Transnational review of visual standards for driving: How Australia compares with the rest of the world. *Clinical & Experimental Ophthalmology, 47*(7), 847–863. https://doi.org/10.1111/ceo.13572

Open Practices Statement The data and materials are available at https://osf.io/tgzb7 and the experiment was pre-registered (https://osf.io/52zes).