A comparison of cybersickness symptoms across 360-degree hazard perception and hazard prediction tests for drivers

Thomas Goodge, Victoria Kroll, Mike Vernon, Petya Ventsislavova & David Crundall

Department of Psychology

Nottingham Trent University

This article was accepted for publication in Applied Ergonomics on Aug 3, 2021. At the time of writing, this article is considered to be 'in press'. Should you wish to cite this article, please update the following reference with the journal volume number and pages numbers once they become available:

Goodge, T., Kroll, V., Vernon, M., Ventsislavova, P., and Crundall, D. (2021). A comparison of cybersickness symptoms across 360-degree hazard perception and hazard prediction tests for drivers. *Applied Ergonomics, [volume number],* [page numbers].

Address for Correspondence:

Prof. David Crundall

Department of Psychology

Nottingham Trent University

Nottingham NG1 4FQ, UK

Email: david.crundall@ntu.ac.uk

Keywords: Virtual Reality, Cybersickness, Hazard Perception, Hazard Prediction

Acknowledgements

This research was funded by the Road Safety Trust, the RAC Foundation and the UK Driver and Vehicle Standards Agency. These sponsors have not materially influenced the design of this study or the reporting thereof.

A comparison of cybersickness symptoms across 360-degree hazard perception and hazard prediction tests for drivers

ABSTRACT

Hazard perception assessment may benefit from VR-presentation by removing field-of-view restrictions imposed by single-screen tests. One concern is whether VR-induced 'cybersickness' will offset any benefits. Self-reported cybersickness ratings were recorded from 77 participants viewing two variants of a 360-degree hazard test: *hazard perception* and *hazard prediction*. The latter was hypothesised to be particularly susceptible as clips abruptly cut to a probe question at hazard onset. Such sudden occlusions are thought to increase cybersickness. Overall cybersickness levels were low, with only four participants excluded for above-threshold sickness ratings. The remaining participants showed unexpectedly lower symptoms for the hazard prediction test and rated this test format as more comfortable and engaging. These findings mitigate concerns over the use of 360-degree videos in formative hazard assessments, even when clips involve sudden occlusions. Nonetheless, removal of any participants due to cybersickness raises problems for using VR for formal assessments of hazard perception skill.

Keywords: Virtual Reality, Cybersickness, Hazard Perception, Hazard Prediction

A comparison of cybersickness symptoms across 360-degree hazard perception and hazard prediction tests for drivers

1. Introduction

Virtual Reality (VR) has seen a rapid increase in fidelity in recent years (Slater & Sanchez-Vives, 2016). Unfortunately, while a plethora of research papers detail a broad range of VR applications, relatively few assess the benefits VR provides over and above more conventional presentation modes (e.g. presenting video content on a single screen). Of the notable exceptions, Ruddle et al., (1999) compared navigation on a desktop system to head-mounted VR, finding no overall task performance differences, though VR users were more likely to stop and inspect other aspects of the scene, which may have contributed to their ability to better estimate vector distances between waypoints (see also Ruddle and Lessels, 2009). In contrast, Srivastava, Rimzhim, Vijay, Singh and Chandra (2019) found evidence in favour of a desktop navigation task compared to a VR equivalent when measuring subsequent spatial learning. Aoki, Oman, Buckland and Natapoff (2008) also found in favour of a desktop system over a VR system for simulating an escape from a space station. Mental rotation and perspective taking have also been compared in VR and on a desktop monitor (Freina and Canessa, 2015). While certain VR advantages were noted (e.g. the ability to literally view the scene from a different perspective), the authors noted that the more immersive environment could distract from the main task and negative impact learning outcomes. In a more applied scenario, Herrera, Bailenson, Weisz , Ogle and Zaki (2018) found more people willing to sign a petition for affordable housing following a VR perspective-taking task focused on homelessness compared to a desktop comparison.

In a different domain, MacQuarrie and Steed (2017) had participants view videos (e.g. horror, documentary, etc.) in a VR headset or on a single screen. They found participants to report greater enjoyment and spatial awareness in the 360-degree mode. In the years succeeding this research, the technology has advanced even further, with VR systems such as the HTC Vive and the Oculus Rift becoming more readily available to the general public. The most recent study to take advantage of the latest generation of headsets, and the most relevant to our current interests, was conducted by Madigan and Romano (2020) who compared hazard training interventions across VR headsets and single screens finding greater training benefits for the VR modality, though they did not directly compare the use of dynamic video clips across these two presentation methods. Currently, wireless headsets (e.g. the Oculus Quest) are leading the next wave of consumer acceptance. Consequently, there has been greater scope and accessibility for VR technology to be used for practical applications such as education and training.

1.1 Hazard Perception in 360-degrees

One potential application of VR technology is within the road safety field, specifically in regard to hazard perception training and assessment. Hazard perception (HP) refers to the skill of identifying an on-road hazard in sufficient time to avoid a collision. The national UK HP test presents viewers with video clips from the driver's perspective, each containing at least one hazard. Drivers must press a button as quickly as possible to acknowledge the hazard. Typically, safer drivers respond faster to hazards than less-safe drivers. Some researchers argue this to be the most clear-cut, higher-order measure of skill relating to driver safety (e.g. Horswill 2017). Decades of research led the UK Government to introduce a hazard perception test as part of the licensing procedure in 2002 (see Horswill & McKenna, 2004, for a review), with subsequent evidence suggesting that this has reduced certain types of collisions (Wells et al., 2008). The test has however been criticised for several reasons (e.g. Crundall, 2016). For instance, the UK hazard perception test is presented on a single screen, displaying perhaps an arc of 60-degrees of visual angle ahead of the car. Without access to any mirror information, or any opportunity to look outside this forward arc, the variety of useable hazards is limited. It is difficult to have hazards that would normally require a driver to look outside this forward arc (e.g. a bicycle positioned to the driver's nearside, or an overtaking vehicle). We know that wider fields of view, and mirror information, change how drivers respond in such tests (Shahar et al., 2010), thus a move to presenting hazards in a 360-degree environment could evoke more ecologically valid driver behaviour. Furthermore, a 360-degree view raises the possibility of assessing other safety-critical manoeuvres such as reversing, parking, or gap acceptance when pulling out from a side-street onto a main carriageway.

This argument was used by Madigan and Romano (2020) to support their development of RAPT training in VR headsets (Risk Awareness Perceptual Training, e.g. Fisher, Pradhan, Hammel, DeRamus, Noyce and Polletsek, 2003). Their ground-breaking study compared the training benefits of still images presented on a computer to still images presented in a VR headset, and to dynamic videos in VR headsets. While they found a suggestion that the VR videos produced the only improvement in hazard perception skill, this result was inadvertently confounded with the amount of training received. In another recent study, a virtual-reality hazard anticipation and mitigation training program (V-RAPT) was shown to improve novice drivers' hazard anticipation skills and their ability to mitigate potential threats to a greater extent than the traditional RAPT training (single-screen) when evaluated on a driving simulator (Agrawal et al., 2018). The traditional RAPT training intervention is not, however, described in the paper. Instead Agrawal et al. (2018) reference a much earlier study to describe their RAPT training condition (Pradhan et al., 2006). A comparison of both papers reveals considerable differences between the two training interventions that confound the conclusion that the V-RAPT superiority is solely due to presenting the training within a VR environment. Despite the confounds noted in the above studies, the findings suggest that there is potential for hazard assessment and training via VR headsets.

1.2 Cybersickness in immersive environments

Though there are practical and theoretical arguments for creating a more immersive hazard perception testing and training environment, any move into the virtual world brings with it a possibility of inducing feelings of nausea and sickness in participants. Within the driving research literature, we acknowledge the existence of *simulator sickness,* which can produce some severe symptoms in a small percentage of participants when using driving simulators (Kolasinski, 1995, O'Brooks et al 2010). Sickness experienced in VR has been described as a distinct phenomenon to simulator sickness, however. Often termed '*VR sickness'* or '*cybersickness',* common symptoms can include nausea, dizziness, and headaches, amongst others, akin to motion sickness or simulator sickness, but with greater severity (Stanney, Kennedy and Drexler, 1997). Disorientation, nausea and blurred vision are specifically exacerbated in virtual reality studies (Kennedy et al., 2003; Kim et al 2018), which has been argued to distinguish the condition from other forms of laboratory-based motion sickness. Regardless of the distinction between cybersickness and general simulator sickness, both are often measured on the same scale (Rebenitsch and Owen, 2016). Perhaps the most popular measure of sickness in simulators and VR is the Simulator Sickness Questionnaire (SSQ; Kennedy et al., 1993).

Despite the potential benefits of VR as a medium for assessing and training drivers, there has been little research to ascertain whether any benefits outweigh the possibility of cybersickness. Of the few studies that exist, Aykent et al (2016) and Forster et al. (2015) reported no benefit or degradation of driving behaviour in VR, though both studies reported feelings of sickness in participants. More recently, Weidner et al. (2017) compared a head-mounted VR driving simulation to a 2D screen version. Presentation mode did not affect lane-changing behaviour, or physiological responses, but the VR once again induced more sickness. More promisingly however, one recent study demonstrated that experienced drivers had better hazard anticipation of latent hazards than novice drivers in a VR test (Mangalore et al, 2019) with less than 10% of participants removed through sickness, which they claimed would not be a problem for most studies. When they compared sickness levels to the same test provided via a simulator, they observed no significant difference between the sickness scores across the platforms. Similarly, the recent hazard training benefits noted by Agrawal et al. (2018) comparing single-screen training to a VR presentation, were accompanied by relatively low levels of sickness, with only one out of 36 participants removed due to excessive symptoms. Given the limited and mixed nature of the level of cybersickness in driving-related VR applications, more research appears necessary.

One theory underlying such sickness is that of *sensory conflict*. This theory states that the mismatch between oculomotor and vestibular sensory input creates these feelings of nausea and disorientation (Bos, Bles, and Groen, 2008). For instance, in a 360-degree driving video, the viewer can visually perceive the movement of the car but does not perceive the associated physical movement in the vestibular system, potentially resulting in sickness. Previous research has shown that the disconnect between perceived experience (in this case visual) and actual experience (e.g. vestibular) is one of the largest factors in experiencing cybersickness symptoms (Weech, et al., 2019). This is supported by Duh et al. (2004b) who found greater levels of reported sickness with increased visual-vestibular conflict, Whitney et al. (2013) who found that people with vestibular disorders report increased levels of motion sickness, and Ng, Chan & Lau (2020) who demonstrated reduced levels of cybersickness when visual and vestibular sensations were synchronised. Given that visually-perceived motion occurs in a typical hazard clip without any vestibular sensation, it is possible that such clips will evoke cybersickness in viewers.

This potential problem is reinforced by the finding that real world carsickness predominantly affects passengers rather than drivers (Rolnick and Lubrow, 1991), which is more likely to occur in safety-critical situations involving harsh acceleration, braking or swerving (Schmidt et al., 2020). Typical hazard perception tests require participants to watch clips without interacting with them. While they are instructed to watch these clips 'as if the driver', the experience is perhaps closer to watching the road ahead as a passenger.

1.3 The influence of age on cybersickness

The literature is not clear as to the relationship between cybersickness and age. Young people have typically been thought to be more susceptible to motion sickness (Paillard et al 2013) whereas older people are less susceptible (Davis, Nesbitt and Nalivaiko 2014). Reason and Brand's (1975) seminal work on motion sickness found that susceptibility was greatest in ages 2-12 and declined with age. In contrast, studies specifically looking at *cybersickness* ratings in adults have found that older groups suffer more than younger (Arns & Cerney 2005, Knight & Arns 2006).

In the driving context, middle-aged drivers tend to show higher severity of sickness scores in VR and simulators relative to the younger drivers (Brooks et al., 2010; Keshavarz et al., 2018). One possible explanation is that younger drivers are likely to have engaged more with VR and simulators through gaming, and that this repeated exposure attenuates cybersickness symptoms. This negative relationship was noted by Knight and Arns (2006), though later studies that have recorded participants' self-reported gaming experience (Gamito et al., 2008), or have manipulated gaming experience via a training protocol (Gamito et al., 2010) have found little or no relationship with cybersickness.

A recent meta-analysis of VR studies suggested lower reported sickness in older populations, (Saredakis et al., 2020), yet this is only based on a small number of studies that included older populations. We argue that given the variety of findings, the influence of age on driving-related cybersickness remains unclear.

1.4 Hazard Perception or Hazard Prediction?

Beyond the problems caused by the use of single screens in typical hazard tests, one other limitation of the traditional hazard perception test, as used by the UK Government, is that it relies on speeded responses to hazards as a measure of skill. A response to a hazard is considered a hit if it falls within a temporal scoring window. Unfortunately the selection of these scoring windows is difficult and can induce both Type 1 and Type 2 errors depending on whether the window is too long (which may then allow responses to events other than the *a priori* hazard) or too short (whereby genuine responses to early clues about the upcoming hazard may fall just before the scoring window).

Fortunately, researchers have trialled over 100 different measures of hazard perception skill from which one could choose an alternative (Moran et al., 2019). For instance, the spatial response windows employed by Wetton et al. (2011) combine a measure of accuracy with response times, producing considerable effect sizes when differentiating between safe and less-safe drivers.

Our preferred alternative testing format is the *hazard prediction* test (Crundall 2016). Rather than relying on speeded response times to hazards, it measures the ability to predict potential hazards by suddenly occluding the clip at the point of hazard development, after which participants are asked "What happens next?". The most recent iteration of the hazard prediction test then presents drivers with multiple-choice options from which to choose (Ventsislavova & Crundall, 2018). The *hazard prediction* test has proved better suited than the traditional *hazard perception* methodology, as used in the UK national test, at differentiating between groups of novice and experienced drivers (Ventsislavova & Crundall 2018) and high-risk and low-risk professional drivers (Crundall & Kroll, 2018). Both studies used naturalistic video clips, though a recent comparison of hazard prediction and hazard perception tests using CGI clips found both tests to be equally effective (Crundall, van Loon, Baguley & Kroll, 2020). Recent evidence has also associated better hazard prediction performance with fewer self-reported crashes in average drivers (Horswill, Hill & Taylor, 2020). Despite the great promise shown by the hazard prediction test, we do not know how well this will translate into a 360-degree format. It may even be more likely to evoke cybersickness than the traditional hazard perception test, as sudden changes in motion, such as those caused by the abrupt occlusions in the prediction test, are not recommended in the development of 360-degree films precisely for the avoidance of cybersickness (Bonato et al 2008).

1.5 The current study

To assess the practical potential of 360-degree hazard tests the current study aimed to investigate a number of questions. First, we wished to assess the overall levels of cybersickness evoked by hazard tests presented in VR headsets. Based on previously reported levels of sickness in VR studies in other domains, we might expect between 20 and 30 percent of participants to be so nauseous that they must withdraw from the study (Stanney et al 1999, Chen et al 2011), although there is contention about this figure. Secondly, we were keen to identify whether age influences cybersickness in such tests. This is an important issue when considering accessibility of training and assessment tools by a wide target audience. Thirdly, we aimed to compare the two hazard test variants when presented in a 360-degree format (perception vs. prediction). Although there are strong evidential, theoretical and pragmatic arguments for favouring a prediction test over a perception test in terms of the behavioural measure (Crundall, 2016, Pradhan & Crundall, 2017), it is a possibility that the abrupt occlusions that are included in the prediction test may result in greater sickness, and this in turn may affect participants’ ratings of perceived comfort, realism, engagement, and immersion. Finally, we had the opportunity to compare participants’ preferences regarding both of these tests.

2. Method

2.1 Participants

Seventy-seven drivers were recruited, split across five different age groups (17-25, 26-35, 36-45, 46-55, 56+). These categories were identified through discussion with stakeholders and ensured a more representative sample of the general population. One participant was removed due to equipment failure. The demographic details of the remaining 76 participants are given in Table 1. One driver was still learning to drive (age 26-35), and one driver had only passed the driving test within 12 months of taking part in the study (and would normally be considered a novice driver, age 17-25). None of the participants reported any ocular problems, and all had normal, or corrected to normal eyesight. Participants were permitted to wear spectacles within the headset where required.

2.2 Design

A mixed 2 x 5 x 5 design was employed with test type as a within-subjects variable (hazard perception vs. hazard prediction). Participant age group was a between-subjects variable (with 5 levels, see Table 1). The third variable was the 5 points in time when cybersickness ratings were provided by the participants. Cybersickness ratings were the primary dependent variable (measured by the 16-item SSQ) and they were collected at three points during the acclimatisation (baseline, after a short practice and once more after a longer practice) and also following both the hazard perception and prediction tests. Ratings were also collected on 10-point Likert scales after both tests for perceived comfort while undertaking the test, perceived realism of the stimuli, the amount of immersion felt within the scene, and how engaged they felt with the task (with higher ratings reflecting greater comfort, realism, immersion and engagement).

Table 1. Demographics for all participants who completed the study

Group	N	Sex	Age (years)		Driving Experience (years since passing test)	
			M	*SD*	*M*	*SD*
17 – 25	22	19 female	20.9	1.6	2.5	1.4
26 – 35	12	6 female	30.9	3.5	10.3	4.7
36 - 45	13	9 female	40.4	2.8	16	6
46 – 55	11	7 female	51.2	3.3	29.4	5
56 +	18	7 female	69.6	7.9	48.4	9.6

Twelve clips were used to create two hazard perception tests containing six clips each. The same 12 clips were then edited to create two prediction tests. Participants either saw hazards 1-6 in a perception format and hazards 7-12 in a prediction format, or vice versa. The order in which the tests were presented was counterbalanced across participants. Limitations on the presentation software (Tobii Pro Lab) meant that we could not randomise the order of clips within each test. Accordingly, two separate orders of clips were created (one sequential order and one reversed order) for each test, and these test variants were factored into the counterbalancing schedule.

Behavioural responses to the clips (hazard perception and hazard prediction scores) and eye movements were collected but, given the small number of clips, these measures were not considered of primary importance to this study which was designed specifically to address issues of cybersickness. These measures are not reported in the current paper.

2.3 Stimuli & Apparatus

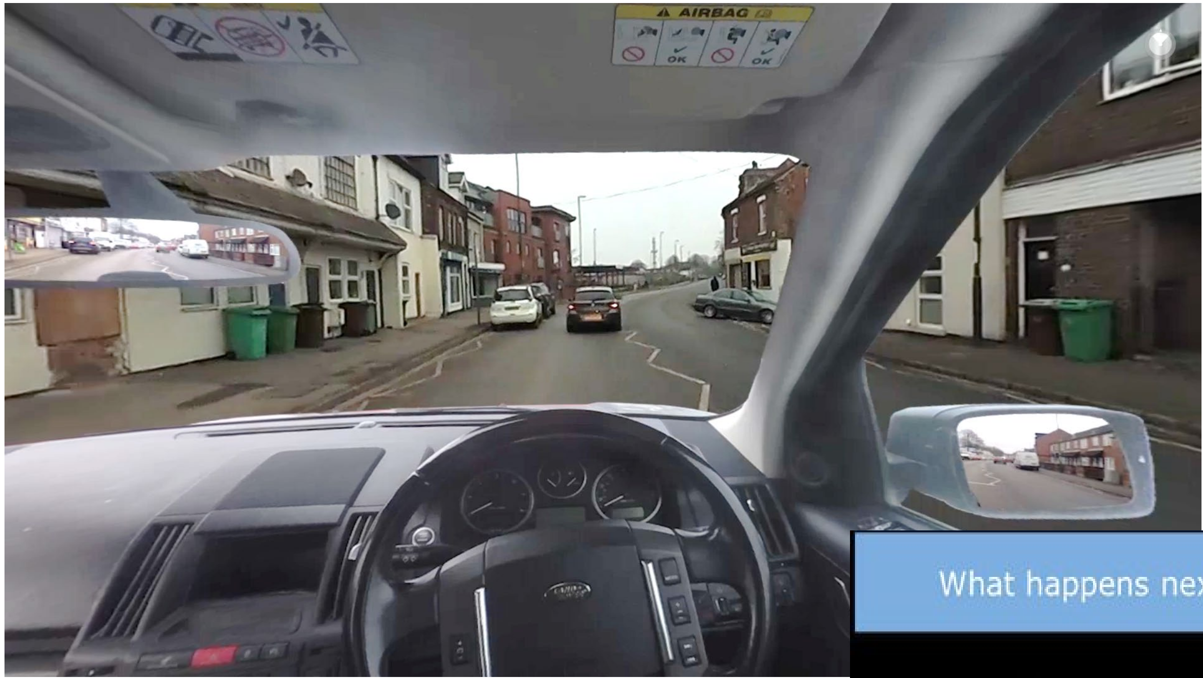
2.3.1 Filming & Editing:

The 360-degree tests were developed from footage recorded specifically for this study from a Garmin VIRB 360 action camera mounted on the top part of the windscreen of a Vauxhall Corsa. This was positioned directly above the driver, to capture the driver's perspective. For mirror information, three GoPro HERO 4 cameras (1080p, 16:9 ratio, medium angle setting) were mounted externally using suction mounts aligned with the mirrors and positioned to avoid visual obstruction for the driver. Twenty hours of driving were undertaken to collect footage of naturally occurring hazards in and around Nottingham city, UK.

A team of traffic psychologists reviewed the footage and selected 12 hazards, based on the following five principles:

(a) There was some danger present that might require the driver to change behaviour (e.g. speed, lane position) to mitigate the possibility of a collision.
(b) The hazard had a precursor (i.e., a clue to the upcoming hazard, e.g. a car approaching the give way line in a side road ahead).
(c) The hazard had an easily defined onset (e.g., a car crossing the give way line of a side road and pulling out in front of the film car)
(d) The hazard should be caused by other road users (rather than caused by the behaviour of our driver who was filming the route)
(e) The driving scene immediately prior to the hazard had to contain other plausible causes of a hazard to allow distracter options to be generated for the hazard prediction version of the test.

Once the hazardous scenarios were selected, the multiple video feeds were synchronised and combined with a graphic overlay of an interior of a Land Rover Freelander. The footage from the Go-Pro cameras was edited into the mirror placeholders, while the 360-degree footage was wrapped around the car interior. This resulted in an immersive video from the perspective of the driver's seat (see Figure 1; panel A). The twelve hazardous scenarios were edited for both hazard perception and hazard prediction formats.



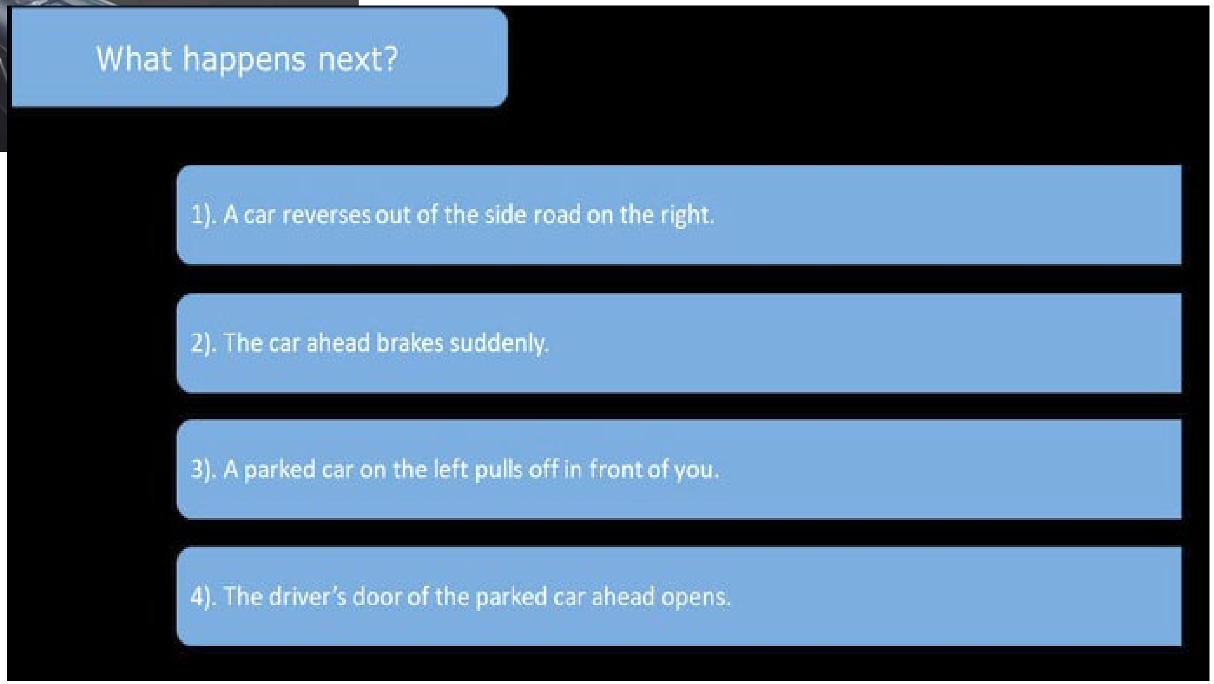


Figure 1. The top panel represents the participants' view from within the VR Headset of a hazard at an occlusion point in the hazard prediction test. The 360 footage was overlaid with an image of the interior of a Land Rover. This overlay image was taken from the point of view of the driver to simulate the view of the road when sat in the driver's seat. The bottom panel shows the multiple-choice option responses for this clip that follow the occlusion point. Option 1 is the correct answer.

2.3.2 The practice clip

Prior to taking part in the study, participants were acclimatised to the VR-headset via a practice clip. This clip was a 360-degree road safety video produced by Road Safety Scotland and was edited to 2 minutes and 13 seconds in length (https://www.youtube.com/watch?v=hnWgEGVjlak). This clip showed the driver's perspective of a test drive (complete with car salesman in the passenger seat). It was used with permission. Participants were instructed to watch the clip as if they were the driver, but they were not required to give any form of response to any observed hazards.

2.3.3 The hazard tests

There were twelve clips in total, each containing an *a priori* hazard, though participants only ever saw six clips in the hazard perception format. The average clip length was 38.7 seconds (range: 24 to 57 seconds) and the mean hazard onset was 25.6 seconds (range: 15.2 to 45 seconds). Hazard onsets were determined by our team of traffic psychologists, who selected the points at which a *precursor* first became visible, and when each precursor turned into a developing hazard. A precursor is defined as a clue to a potential upcoming hazard, though the outcome of the event is not yet certain. For example, when we first see a car approaching the main carriageway from a side road, this is considered to be the onset of the precursor to a potential hazard, but when it is clear that this vehicle will pull out in front of you, it then develops into a hazard. The hazard onset may be easy to define in some instances (e.g. when the approaching car crosses the give-way line of the side road), though often it is more difficult (e.g. if the car in the side road approaches the main carriageway at high speed, then the point at which the hazard becomes inevitable may occur prior to it reaching the give-way line).

In addition to the precursor onset and hazard onset points, the team identified a hazard offset (the point at which any response would be too late to mitigate the danger). The hazard onsets and offsets defined the temporal scoring windows that participants had to respond within to score points. Following the scoring system used in the official UK hazard perception test, the scoring window was divided into 5 equally sized sub-windows. A response in the first sub-window scored 5 points. A response in the second sub-window scored 4 points, and so on until the response fell outside the scoring window and scored zero points. Any responses that fell before the start of the scoring window were also awarded zero points.

The twelve hazard perception clips were further edited to produce hazard prediction clips. Each clip was occluded immediately following hazard onset (as defined in the hazard perception test, see Table 2). Hazard prediction clips were approximately 12 seconds shorter that the hazard perception clips, typically occluding just after the point of hazard onset. Following the occlusion, four text options were edited to appear in the VR display providing alternative predictions of what might happen next (see Figure 1, bottom panel). The process for developing these options has been described elsewhere (e.g. Ventsislavova and Crundall 2018). Selection of the correct answer scored one point, with a total of six points available across a hazard prediction test.

Table 2. A description of the 12 hazards used to create the two tests (with correct answers underlined).

Hazard No.	Description	Multiple-choice options	Full Clip Length (secs)	Hazard Onset (secs)	Hazard Offset (secs)	Occlusion point (secs)
1	While driving along a suburban route, a white car suddenly appears in a side road to the left, and then pulls out in front of you.	An oncoming car encroaches on your lane. A car pulls out in front of you from the side road on the left. A pedestrian steps out into the road from the left. A pedestrian runs out into the road from the bus stop on the right.	51	41.8	45.1	42
2	While travelling along a busy road with many parked cars, a grey car starts to reverse out of a side road on the right as you approach it.	A car reverses out of the side road on the right. The car ahead brakes suddenly. A parked car on the left pulls off in front of you. The driver’s door of the parked car ahead opens.	54	35.5	40.8	38.9
3	You are driving along a two-lane road, when a car stopped the left lane ahead causes you to change lanes to avoid a collision.	An oncoming car turns across your path into the side road on the left A cyclist emerges from the side road on the left. A parked car blocks your lane ahead. A car overtakes you from the right.	29	24.3	27.9	25.7
4	You are travelling along a residential road, approaching a pair of traffic lights on green. Before you reach them, a white van pulls out in front of you from a side road on the right.	A pedestrian steps into the road from the left. A van pulls out in front of you from the side road on the right. A parked car on the left pulls off in front of you. The car ahead brakes suddenly for the pedestrian crossing ahead.	37	23.3	27.4	23.5
5	Travelling down a busy town centre high street, a car pulls up on the left blocking your path. As you slow down, two pedestrians use this to	Pedestrians step out into the road from the right. A parked car on the left pulls off in front of you. The driver’s door of the car ahead suddenly opens.	43	25.2	33.3	25.3

	cross the road in front of you.	<u>Pedestrians step out into the road from the left.</u>				
6	You are driving a long a main road with a time-restricted bus lane on your left. In your rear-view mirror, you can see a car approach you at speed, which then undertakes you by using the bus lane.	A car emerges from the side road on the left. A pedestrian steps from the right into the road at the crossing. An oncoming car turns across your path into a side road on the left. <u>A car undertakes you at speed.</u>	57	45.2	47.9	45.8
7	You are driving along a main arterial road approaching a pedestrian crossing. Before you reach it, a pedestrian not at the crossing steps into the road from the left	The van in the adjacent lane to the right pulls in front of you. A pedestrian crosses the road from the right. <u>A pedestrian steps out into the road from the left.</u> A car pulls out from the left, blocking your path.	37	22.8	27.9	23.2
8	In a residential area, the traffic lights you are stopped at turn green and you turn left. As you turn, roadwork signs are visible on the left and a road worker steps into the road from the right.	A parked van pulls into the road from the right blocking your path. <u>A road-worker steps into the road from the right.</u> An oncoming van approaches in your lane A fallen road sign blocks your lane.	38	22.1	27.3	23.7
9	You are queueing in slow moving traffic on a two-lane road. As it starts to move, a car from a side road on the right moves to enter the road.	The car in the right lane indicates and pulls into your lane. <u>A car pulls out from the side road on the right.</u> A pedestrian gets out of the car ahead A motorcyclist overtakes you from the right.	24	16.4	25.0	13
10	You are driving at speed on a two-lane arterial road. As you approach a bend, a parked lorry with its hazard lights on becomes visible, and you have to change lanes to avoid a collision.	A pedestrian steps into the road from the left. <u>A parked lorry blocks your lane.</u> A white van overtakes you from the right. A car pulls out in front of you from the side road on the left.	28	15.9	21.9	19.8
11	You are driving along a heavily congested road. A delivery van ahead	A pedestrian steps out into the road from the left	27	15.2	18.1	15.6

	indicates and begins to turn into a side road on the left but must stop and block your path for some pedestrians crossing the road.	A pedestrian with a pushchair appears in the road from behind the turning van. The van ahead brakes suddenly to avoid pedestrians stepping out into the side road. The parked pulls into the road from the left blocking your path.				
12	You drive across a traffic light-controlled crossroads with parking spaces ahead on the left side of the road. The car in front stops, and then reverses to enter one of these spaces, causing you to stop and manoeuvre around it.	A pedestrian steps out from behind the parked car on the left The car ahead performs a U-turn in the road. The white car parked on the left pulls out in front of you. The car ahead reverses towards you.	39	18.9	27.0	24.8

2.3.4 Questionnaires

All participants who completed the study were asked to complete a driving history questionnaire with 11 questions probing annual mileages, weekly hours driving, post-license experience in years, collisions, and license endorsements.

A Simulator Sickness Questionnaire (SSQ) was administered at five points during the testing session. Participants had to rate the severity of 16 symptoms on a four-point scale (from 'none' to 'severe'). Symptoms included sweating, nausea, fatigue etc. (Kennedy et al., 1993). Testing was aborted if a participant reported any single symptom as 'severe', if any symptom rating increased by 2 levels during the testing session, or if three or more factors increased by 1 level during the testing session.

Following exposure to both the hazard prediction and hazard perception tests, participants were given four ratings to complete on a 10-point scale (which we refer to hereafter as the CRIE questions):

a) *Comfort* - How comfortable did you feel while undertaking the test in the VR-headset?
b) *Realism* - How realistic were the video clips (e.g. how close to real life was it)?
c) *Immersion* - How immersive was the test (e.g. did it feel like you were there)?
d) *Engagement* - How engaged did you feel with the task (e.g. were you motivated to continue)?

2.3.5 Apparatus

Both tests were presented in an HTC Vive headset running Tobii Pro Lab, controlled via an ASUS Republic of Gamers Strix Hero III Gaming Laptop. The HTC Vive has a resolution of 2160 x 1200 (1080 per eye) and a 110-degree Field of View (FOV). In the hazard perception test, participants responded via a keyboard. For the hazard prediction test, participants gave their answer verbally to the researcher who recorded their answer via the laptop keyboard.

2.4 Procedure

The following procedure was reviewed and approved by the School of Social Sciences Research Ethics Committee, based at Nottingham Trent University. The research was conducted in accordance with the Declaration of Helsinki and the ethical guidelines stipulated by the British Psychological Society.

Upon arrival at the laboratory, participants were given written instructions. They then signed a consent form, completed a demographic questionnaire and filled in the baseline cybersickness checklist (the first of five occasions when they would fill in this form). The baseline responses ensured any pre-existing symptoms were captured for later analysis. They were given a brief acclimatisation period to the 360-degree environment, viewing 40 seconds of the practice clip (hereafter referred to as the *short practice*). After a brief break, this was followed by a longer acclimatisation period using the full 2 minutes and 13 seconds of the practice clip (the *long practice*). All participants went through this two-step process (short practice followed by long practice) which was designed to gradually build up resilience to any cybersickness symptoms. The sickness checklist was re-administered after both practice clips to check whether participants responded negatively to the environment.

Following acclimatisation, half the participants viewed six hazard perception clips, while the other half were presented with six hazard prediction clips. Participants then filled in another sickness checklist and gave ratings for the four CRIE questions. The second set of 6 clips was then presented, either as a perception test or a prediction test (whichever they had not seen first). The final step was to fill out a fifth sickness checklist and give their CRIE ratings for the last 6 clips.

All participants were seated on a chair in the centre of the laboratory that had been calibrated for the VR headset. The headset was fitted to their face and head and they were told that they were about to watch video clips that were in 360 degrees, taken from the perspective of the driver.

For the hazard perception test, they were instructed to press the space bar on a keyboard as quickly as possible to indicate the presence of a hazard that would require them to suddenly stop, slow down or change position in some way to avoid a potential collision. The keyboard was placed in their lap and participants rested their hand on the space bar.

For the hazard prediction test, participants were instructed to watch the clips and search for potential hazards. They were told that the clip would suddenly stop, and the image would be occluded just as the hazard begins to develop. Following this they were told to expect four possible options (numbered 1 – 4) regarding what might happen next, from which they had to select the correct answer. They verbally reported the number of their selected option, which the researcher entered via a keyboard. This input method was adopted as participants would not be able to view the keyboard to select their answer, without taking off the VR headset and breaking immersion. Participants were aware that there was no speed stress on their responses in the hazard prediction test.

At the end of each practice and the first experimental block, participants were able to remove the headset for a brief break while the SSQ was administered. Breaks lasted typically between 1 and 2 minutes.

3. Results

Four participants (5.3%; 1 in the 17-25 age group, 2 participants in age group 46 – 55 and 1 for the 56+ group) were removed due to their sickness ratings rising above threshold. The demographics of these drivers are shown in *Table 3.*

Table 3. Demographics for four participants who were removed due to sudden increases in SSQ scores

Group	N	Sex	Age	Driving Experience (years since passing test)
17 – 25	*1*	*1 female*	*21.4 yrs*	*4.5*
46 – 55	*2*	*2 female*	*52.6 & 51.2yrs*	*25 & 30*
56 +	*1*	*1 female*	*58.8 yrs*	*30.8*

3.1 Sickness Ratings over Time

For each time point, each participant's total sickness severity was calculated following the method explained in Kennedy et al. (1993). This requires individual item scores (0, 1, 2, or 3) to be summed within three subscales (nausea, ocular discomfort and disorientation; with some items loading on two subscales) for each participant. These subscale scores are then summed together and multiplied by a constant (3.74) to arrive at a total severity score. This total score can vary from zero to 236. Kennedy et al. (2003) suggested that scores between 10-15 reflect significant sickness symptoms. Scores between 15-20 are a cause for concern, while scores over 20 suggest that there is a problem that will probably prevent the simulator from being used until rectified.

The measures in Figure 2 reflect the scores of the 72 participants who were retained in the study following the long practice (i.e. those who did not exceed our withdrawal criteria on the SSQ). The four participants who were removed provided similar ratings to the whole sample at baseline and after the short practice (with means of 15 and 10, respectively), though the ratings they gave after the long practice jumped considerably to a mean of 70 on the SSQ.

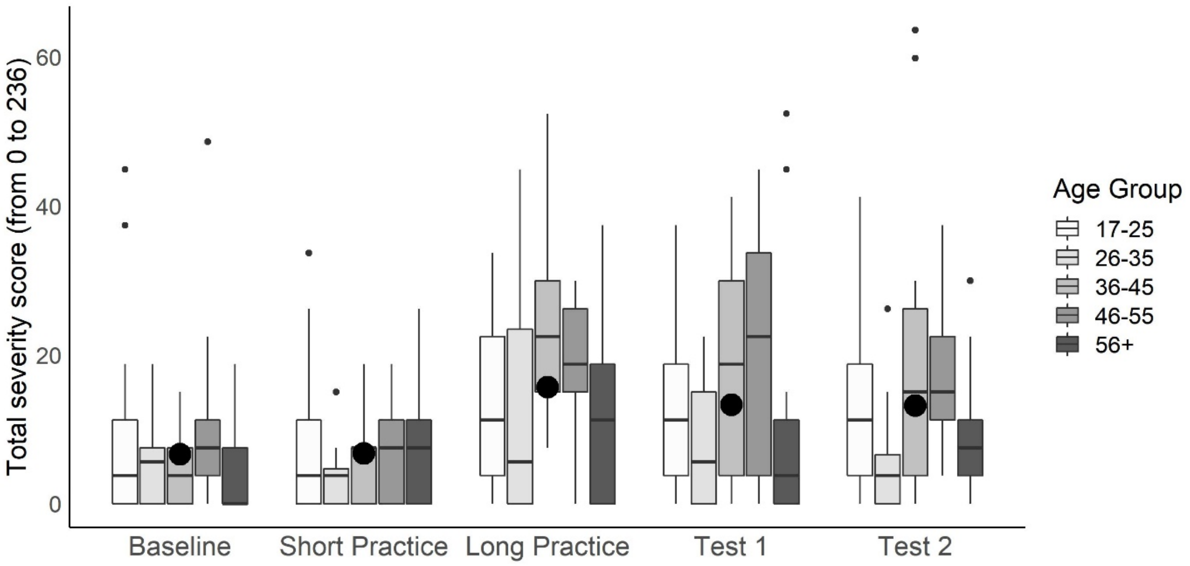


Figure 2. A box plot depicting SSQ ratings, collected at five different points during the study, for the five age groups. The four participants who were removed due to sickness are not included. The large black dots represent the mean SSQ scores for the five points in time at which they were measured.

Mean scores were subjected to a 5 x 5 mixed Analysis of Variance (ANOVA; time of questionnaire x age group, not including the 4 participants who were removed for sickness). The time at which the questionnaire was administered produced a significant effect ($F(4,268) = 17.7$, $MSe = 70.1$, $p < .001$, $\eta_p^2 = .21$). Repeated contrasts across the time factor (baseline to short practice, short practice to long practice, etc.) revealed that the only difference between the times was a significant increase in sickness ratings following the long practice, compared to the ratings given after the short practice (even after Greenhouse-Geisser and Bonferroni corrections: $F(1,67) = 40.1$, $MSe = 150.0$, $p < .001$, $\eta_p^2 = .38$). It appears the *short practice* (40 s) may not have been long enough to raise sickness ratings above baseline, though the subsequent *long practice* resulted in a significant jump in this measure (even without the four removed participants contributing to the analysis). Neither the main effect of age nor the interaction between age and the time when the SSQ was administered were significant.

3.2 Sickness Ratings for Hazard Prediction and Perception

The data in Figure 2 give the sickness scores for the two tests that participants undertook in the temporal order that they took them. For instance, 'Test 1' includes sickness scores for half the participants who saw the prediction test and half the participants who saw the perception test. The scores that make up 'Test 1' and 'Test 2' were recategorised according to their test format (either hazard perception or hazard prediction) and were compared via a mixed 2 x 5 ANOVA (test type x age group).

Mean ratings were entered into a mixed 2 x 5 ANOVA (test type x age group). This analysis revealed a main effect of test type, $F(1, 67) = 5.18$, $MSe = 44.66$, $p = .03$, $\eta_p^2 = .08$, with participants in the hazard perception test reporting higher sickness scores than those in the hazard prediction test ($M = 14.9$ versus $M = 12.3$, respectively). While the absolute difference between the two test is slight, this significant result in favour of the prediction test is sufficient to conclude that it did not produce greater sickness symptoms than the hazard perception test, as initially suggested.

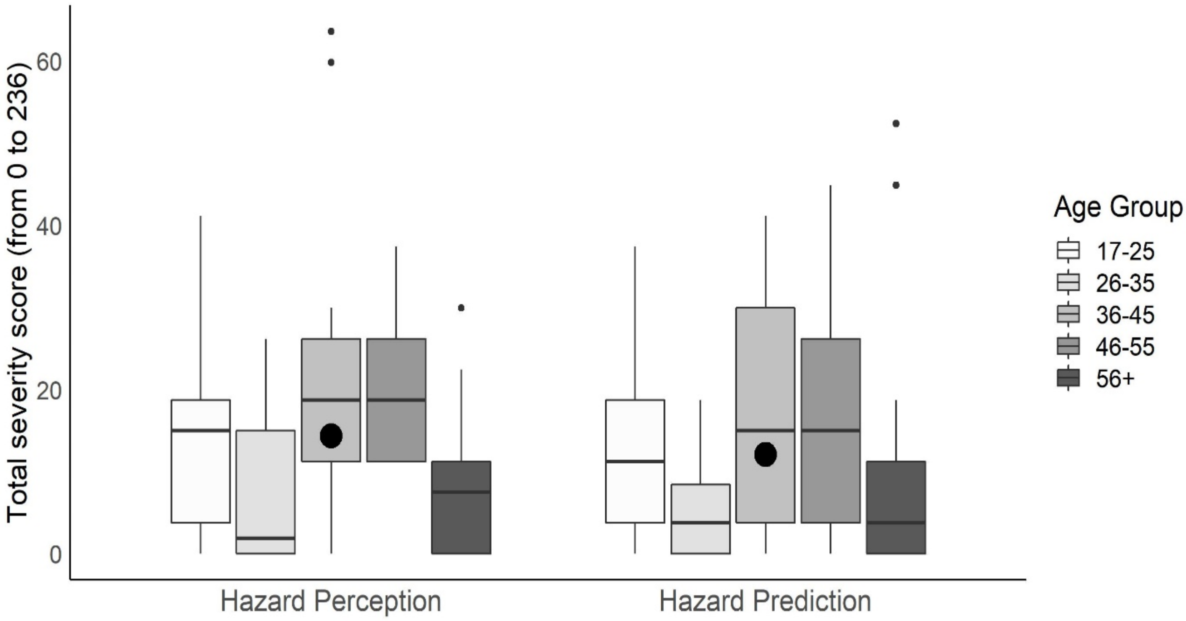


Figure 3. Boxplot depicting sickness ratings for the hazard perception and prediction tests.

3.3 Comfort, Realism, Immersion and Engagement Questionnaire (CRIE)

Participants gave ratings for each of the questions regarding comfort, realism, immersion and engagement on a 10-point scale for both the hazard perception and the hazard prediction tests. Each rating was entered into a within-subjects 2 x 5 ANOVA (test type x age group).

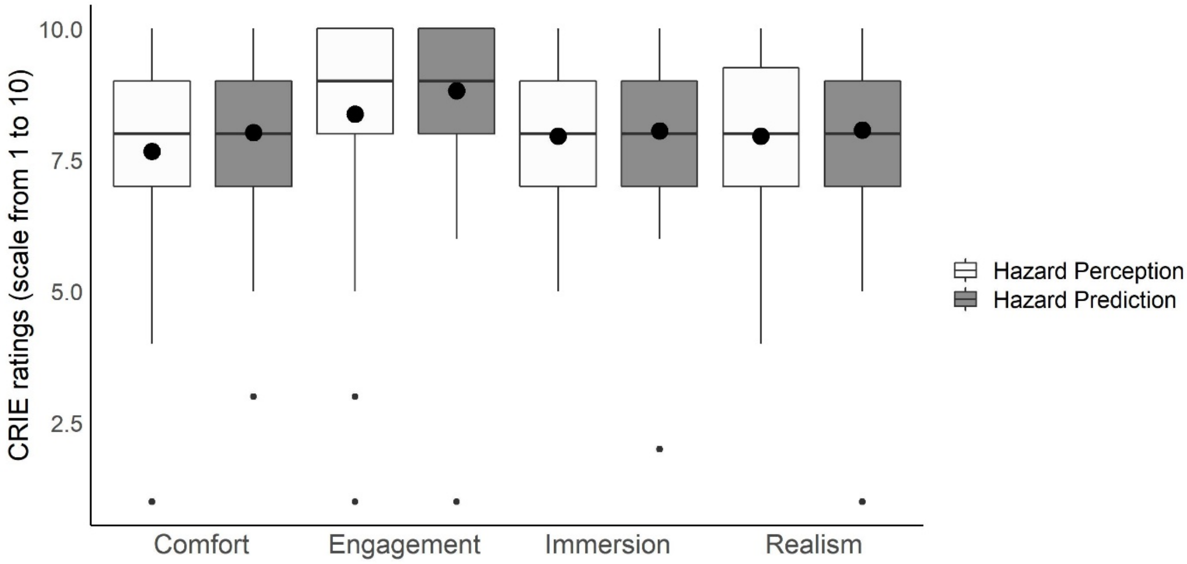


Figure 4. Boxplot depicting the CRIE ratings, combined across age groups, for the two test types.

Main effects of test type were found for comfort and engagement ratings, with participants rating the hazard prediction test as the more comfortable ($F(1, 67) = 6.4$, $MSe = .74$, $p = .01$, $\eta p^2 = .08$, with mean scores of 8.0 and 7.6) and more engaging of the two tests ($F(1, 67) = 10.73$, $MSe = .76$, $p = .002$, $\eta p^2 = 0.14$; with means of 8.8 and 8.3). The means for all CRIE questions are displayed in Figure 4.

Comfort ratings were also affected by age ($F(4, 67)= 2.8$, $MSe = 3.9$, $p = .03$, $\eta p^2 = .14$), suggesting that participants aged 46-55 (M = 7.11) and 56+ (M = 7.4) found the tests to be less comfortable than those participants in the 26-35 (M = 8.9) age group (Figure 5). The interaction between test type and age group approached significance for comfort ratings, but fell short of the threshold, $F(1, 67)= 2.4$, $MSe = 1.76$, $p = .059$, $\eta p^2 = .13$. Figure 5 suggests that the better comfort ratings of the prediction test were primarily reported by those participants over 35. There were no other significant main effects or interactions for any of the CRIE questions (all *p*'s > .05). Correlations between the CRIE ratings and age, supported the conclusion that older participants reported less comfort, at least in the hazard perception test (see Table 4 for the correlations). The correlations also suggest stronger associations between engagement and the other three factors of comfort, realism and immersion.

Table 4. Pearson correlations between CRIE ratings and age for the two hazard tests.

		Hazard Perception			
		C	R	I	E
	Age	$-.291^{*}$	-0.032	-0.004	-0.097
Hazard Perception	C		0.110	$.301^{*}$	$.392^{**}$
	R			$.718^{**}$	$.248^{*}$
	I				$.274^{*}$
	E				

		Hazard Prediction			
		C	R	I	E
	Age	-0.095	-0.043	-0.079	-0.139
Hazard Prediction	C		$.436^{**}$	$.438^{**}$	$.502^{**}$
	R			$.737^{**}$	$.633^{**}$
	I				$.658^{**}$
	E				

Following the correlations, hierarchical regressions were undertaken to assess the impact of age and CRIE ratings on sickness scores for the two tests (on the assumption that sickness is the logical criterion variable). Participant age was entered at stage one of each regression, followed by CRIE ratings at stage 2 (see Table 5). The regression model for *hazard perception* sickness revealed that at stage one, participant age did not contribute significantly to the regression model, $F(1, 70) = .23$, $p = .64$, and only accounted for 0.3% of the variance in hazard perception sickness. Adding stage 2 to the regression model accounted for 12.6% (ΔR^2 = 12.3%) of variation in hazard perception sickness, but this did not reach the threshold for significance, $F(5, 70) = 1.9$, $p = .110$.

When this analysis was repeated for the *hazard prediction* sickness scores, age did not predict sickness, $F(1, 70) = .08$, $p = 0.77$, and only accounted for 1% of the variance in sickness scores. However, adding the CRIE ratings resulted in a model that accounted for 15.3% (ΔR^2 = 15.2%) of variation in sickness scores. This change in R^2 approached the threshold of significance, $F(5, 70) = 2.35$, $p = .05$. However, of the four CRIE questions, the only factors that influenced hazard prediction sickness were comfort ($\beta = -.43$, $t(65) = - 3.18$, $p = .002$) and engagement ($\beta = 0.37$, $t(65) = 2.10$, $p = .04$).

Table 5. Factor relationships with cybersickness ratings.

	Dependent variable	Independent variable	Unstandardized coefficients (B, standard error)		Standardised coefficients (β)	t-value	R^2 (adjusted R^2)	F-Value
Hazard Perception	Model 1	Age	-0.41	0.09	-0.06	-0.47	0.003 (-0.01)	0.64
	Model 2	Age	-0.12	0.09	-0.161	-1.32	0.13 (0.06)	0.11
		Comfort	-3.01	1.11	-0.37	-2.71		
		Realism	-1.81	1.43	-0.22	-1.27		
		Immersion	0.97	1.78	0.10	0.54		
		Engagement	0.95	1.10	0.11	0.87		
Hazard Prediction	Model 1	Age	.02	.08	.04	.292	0.001 (-0.01)	.09
	Model 2	Age	0.03	0.08	0.05	0.389	0.15 (.09)	2.35*
		Comfort	-3.98	1.25	-.432	-3.18*		
		Realism	0.33	1.47	-.04	-.22		
		Immersion	0.19	1.62	0.2	0.12		
		Engagement	3.47	1.64	0.37	2.10*		

Note: For *hazard perception* step 1, $\Delta R^2 = 0.003$; for step 2 $\Delta R^2 = 0.123$. For *hazard prediction* step 1, $\Delta R^2 = 0.001$; for step 2 $\Delta R^2 = 0.151$; *$p \leq 0.05$.

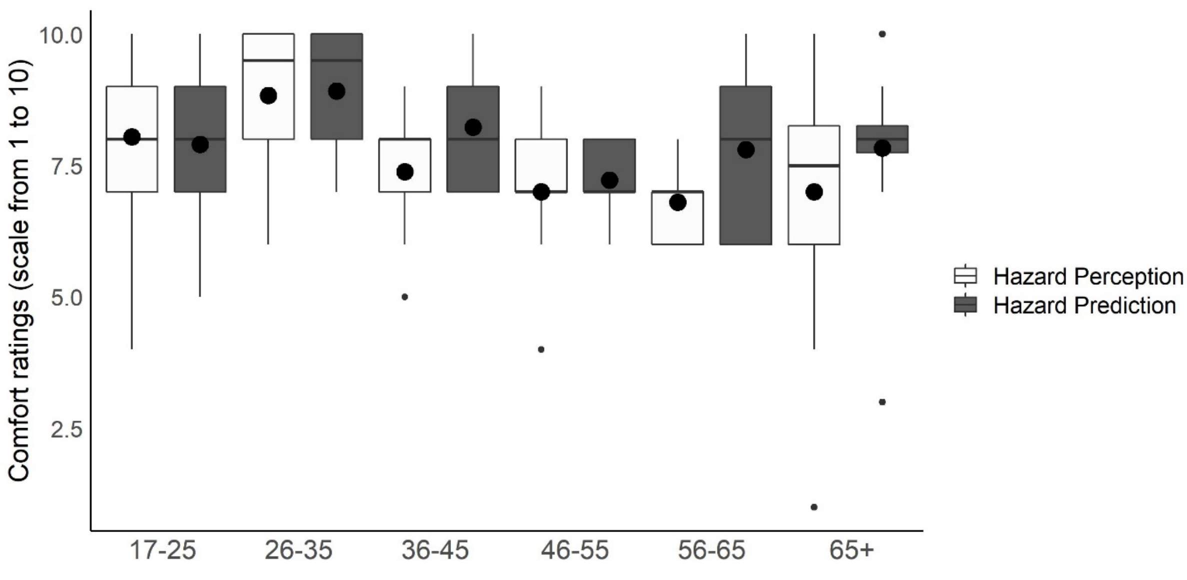


Figure 5. Boxplot depicting the Comfort ratings of participants across the tests and age groups. The large black dots denote the condition means.

4. Discussion

This study was undertaken to assess the extent to which 360-degree hazard tests, using recorded footage from a moving vehicle, might induce cybersickness. The experience of the four participants who were removed from the study will be discussed first, before moving on to the sickness ratings of the remaining 72 participants. The difference in sickness scores evoked by the two tests will then be addressed.

4.1 The experience of cybersick participants

The four participants who were removed following the long practice averaged a score of 70 on the SSQ. While this does not appear large in respect to the extent of the scale (0-236), their scores are extreme compared to those of the other participants. All four participants were glad to be removed from the study at this point. Two of the participants expressed scepticism regarding cybersickness prior to putting the headset on but were convinced of its reality shortly after. Two participants also reported feelings of guilt at not continuing but were given reassurance by the experimenter that withdrawal was necessary and did not reflect negatively on them.

During debriefing, these participants reported that their symptoms had gradually increased throughout the long practice clip, but the final turn of the film car caused a step-change in their discomfort. This turn can be viewed at 2 minutes 34 secs in the publicly available clip. It involves a slight turn to the right, followed by a sharper turn to the left into the side road, and finishes with a second slight turn to the right following a curve in the road. The approximate change in directional heading is 30 degrees to the right, 75 degrees to the left, and another 30 degrees to the right. The simulated speed going into the first bend was 12 mph, while the exit speed from the final bend was 27 mph. Given the number of variables involved it is difficult to conclude whether the subjective step-change in discomfort was caused by the magnitude of the left turn, the combination of all three turns, or the apparent acceleration during the turns. Nonetheless, such sharp bends are likely best avoided in future VR hazard tests (Wetton et al., 2011). Notably, none of the four participants raised issues with discomfort until the clip had finished and they were asked to report their symptoms.

From these reported experiences, we argue self-identification of susceptibility to cybersickness is unlikely to be a reliable screening tool given the scepticism of some of our cybersick participants (i.e. relying on self-exclusion from VR participation is not sufficient to identify all those participants who may develop symptoms). It also appears to be important to continuously monitor symptoms: Even though participants are briefed that they can withdraw at any point without explanation, they may feel obliged to continue and may not voluntarily report symptoms until a much higher threshold is passed (which may require a prolonged period of recovery). The combination of turns at the end of the practice clip proved particularly effective at identifying these participants, and we recommend that practice/screening clips include similar events, especially if the subsequent test is likely to include similar turns. The length of any practice/screening clips may also be important. At over 2 minutes, this clip was the longest that participants encountered in this study, and this may have contributed to the increased discomfort. In shorter clips, a gradual build-up of discomfort may be ameliorated, or periodically reset, by the gaps between clips. A prolonged exposure may however create a gradual build-up of discomfort that eventually reaches a tipping-point.

Screening-out participants who are likely to present symptoms is perhaps the biggest challenge for using VR in an experimental setting (Davis et al., 2014). While our four cybersick participants were not screened-out prior to putting on the headset, the ability to remove them from the assessment within a few minutes represents a time saving for all involved, and ensures that such participants are not subjected to a longer period within the headset. If these participants had started the first of the two hazard tests, they may have endured a further 5-6 minutes of 360-footage before the next iteration of the SSQ identified a problem. Given that previous studies of sickness in VR and simulators has found that increases in exposure of up to 10-minutes can significantly increase levels of reported symptoms (Min et al., 2004; Moss and Muth, 2011), our two-minute practice clip was very useful in identifying cybersick participants early in the process. We do not consider, however, such early screening to be a fool-proof method of identifying all potentially sick participants. We recommend continual assessment of sickness ratings throughout the subsequent study.

4.2 Sickness ratings for those who remained in the study

The results suggest that the sample reported a significant increase in sickness symptoms following the long practice. Symptoms them plateaued for the two experimental tests. Despite the significant increase in symptoms, the individual reports did not meet the pre-determined criteria for removal of any of the 72 participants who went on to undertake the hazard tests. While there is a clear difference between the ratings of the four participants who were removed from the study and the ratings of our remaining participants, we must still question whether the significant increase in symptoms for the surviving sample represents a problem for these VR tests.

To assess whether the sickness ratings evoked by the hazard tests in the current study are acceptable, we need to compare them to a baseline. Kennedy et al. (2003) suggested that scores from 10 to 15 reflected significant symptoms. From 15 to 20, SSQs scores should be of greater concern, while scores above 20 reflect problems that are probably inherent in the simulator rather than in a small number of participants. Given that our participants reported a mean score of 6.6 before even putting on the headset, the probability of remaining below Kennedy et al.'s (2003) level of acceptability was unlikely. Following the long practice, the mean rating for the 72 participants who remained in the study had risen to 15.3, dipping slightly thereafter to 13.1 for both subsequent tests. Following Kennedy et al.'s guidance, these tests still cause significant levels of sickness. Some of our age groups even breach the 20-point barrier to suggest that our tests have inherent problems.

Our 360-degree content does not appear to be alone in the creation of apparently serious levels of symptoms, however. Saredakis et al.'s (2020) meta-analysis of SSQ scores from 55 VR studies found a mean sickness score of 28.0, ranging from 14.3 to 35.2. In addressing the disparity between modern VR sickness scores and the cut-off points given by Kennedy et al. (2003), Saredakis et al. pointed out that Kennedy's scores were based on the self-report of military pilots. We have previously noted problems with relying on self-report in police drivers, where there is the possibility that social desirability might limit self-report ratings despite contrary physiological changes (Crundall et al, 2003). When asking military pilots how sick they feel in a flight simulator, it is possible that the subsequent scores reflect, in part, attempts to manage and maintain a professional image. Furthermore, the flight simulators that Kennedy used are likely to have had a limited level of background detail, compared to that required for ground transport simulation. Rotating in an empty sky may simply be less nauseating than turning a corner on a busy city street.

Instead of comparing our sickness ratings to those of Kennedy et al. (2003), it is perhaps more useful to compare them to the results of Saredakis et al.'s (2020) meta-analysis. The average sickness ratings for our hazard perception test (14.3) falls at the lowest end of the range of scores reported in the 55 studies reviewed by Saredakis et al., while the mean rating for the hazard prediction test was even lower (12.1). These tests did not include ratings from the four participants who were removed following the long practice, but if we assume that their sickness ratings would have plateaued at 70 for both tests, this would give revised sickness ratings of 17.2 and 15.1 for the perception and prediction test, respectively. These sickness scores are still at the lower end of the range reported by Saredakis et al. (2020). Even those age groups that were ostensibly (but not significantly) most affected by cybersickness symptoms (36 – 45, 46 – 55) still gave ratings considerably below the mean sickness ratings found in the meta-analysis (<28). Based on this comparison, it appears that our 360-degree hazard tests evoke remarkably low levels of cybersickness in most participants.

4.3 Explanations for the low rates of sickness

If one accepts the argument that we should compare our ratings to those of Saredakis et al., (2020) rather than to the older guidelines of Kennedy et al., (2003), the results suggest that both tests fare remarkably well, producing sickness ratings at the lower end of expectations within the VR domain. This may be explained by several factors including the brevity of the tests. Each test lasted no longer than 5-6 minutes. Taken together, and including the practice clip, participants were immersed for less than 15 minutes in total and took several breaks from the headset between the practices and the tests. This is much shorter than other studies with higher dropout rates (Saredakis et al., 2020).

Another potential reason behind the lower sickness levels reported could lie in the format of the stimuli presented to participants. Previous research has found that the inclusion of a static independent visual background reduces levels of simulator sickness in driving studies (Duh, Parker & Furness, 2001, 2004a). It has been suggested motion sickness is not caused by conflicting motion cues, but conflict with the so-called ‘rest frames’, which are parts of the virtual environment which are consistent with the real world (Prothero, 1998; Prothero & Parker, 2003). In this instance, the graphic overlay of the car interior could have provided an independent background for participants to orient themselves by, reducing the overall mismatch between the virtual and physical environment.

The low sickness ratings in the current study also mitigate concerns that the passenger effect noted with real-world carsickness is a problem (Rolnick and Lubow, 1991). Whilst this could have potentially been off-set by inducing some form of interaction into a VR test, it is possible that providing the instructor’s voice over (e.g. “Turn left at the traffic lights”), provided sufficient opportunity to anticipate visual shifts in perspective akin to the anticipation provided through interaction.

4.4 Hazard Perception vs. Hazard Prediction

Hazard prediction has previously been found to be a more effective format for measuring hazard awareness skills in studies using naturalistic video clips compared to the traditional method employed in the UK national test, possibly through reducing criterion bias and removing problems associated with early button responses (Crundall 2016; Crundall & Kroll, 2019; Ventsislavova et al., 2019). However, we were concerned that the prediction-test format might lead to greater reported levels of sickness due to the use of sudden occlusions during clips. Typically, mismatch between participants’ perceived and actual movements tends to increase levels of reported sickness (Bos, Bles, & Groen, 2008), with rapid changes in the visual modality producing the highest levels of sickness (Bonato et al., 2008). Thus, sudden occlusions might have increased participants’ discomfort.

Even though both tests induced relatively low sickness ratings, contrary to expectation, the hazard prediction test evoked significantly lower levels of sickness than the hazard perception test. One possible reason for this might be found in the shorter length of the prediction clips compared to the perception clips, with the latter being over 11 seconds longer on average. Across a test of 6 clips, this results in an extra minute within the VR headset. It is possible that this slight difference in test duration was enough to offset the potential detrimental effects of the sudden occlusions in the prediction test. A second explanation may lie in the time that participants spent reading the multiple-choice options in the prediction test, which provided a break between successive video clips. This may have allowed any symptoms arising from the dynamic nature of the stimuli to settle before clips. Given the higher comfort ratings of participants over 35 in the prediction test, they may have benefitted more than younger participants from the shorter clips and breaks provided by the multiple-option screens.

It is also likely that the lower sickness rates in the prediction test may be partly explained by the higher comfort ratings given by some participants. The relationship between comfort and sickness was borne out in a regression, accounting for a modest but significant amount of variance. In addition to being rated significantly more comfortable, the hazard prediction test was also reported to be more engaging than the hazard perception test. A previous study comparing single-screen prediction and perception tests resulted in several participants commenting that the prediction test 'kept them on their toes' and was 'more like real life driving' (Crundall et al., 2021), which could partially explain the effects on engagement in the current study. The regression suggested however that engagement was positively related to sickness symptoms (higher engagement is related to higher levels of cybersickness). Given the relatively low sickness ratings, this relationship between engagement and sickness is not a great concern, though it is puzzling. It is hoped that further research will unpack this effect.

4.5 The influence of age on cybersickness

As noted in the introduction, the existing literature that discusses age and cybersickness is inconsistent (Kennedy et al., 2010; Benoit et al., 2015). There are several researchers who argue that advanced age is a good predictor for simulator sickness (Cassavaugh et al.,2011; Classen et al., 2011; Golding 2006; Trick & Caird, 2011). For instance, a recent survival analysis of a driving simulator resulted in nearly 60% dropout due to simulator sickness from a sample of 88 with a mean age of 73 years (Matas, Nettelbeck and Burns, 2015). Despite this evidence, Saredakis et al. (2020) reported that their meta-analysis suggested cybersickness may be less problematic in older drivers than previously thought, though they caveated this conclusion with concerns over the small number of studies involving older participants.

Unfortunately, the results of the current study do not shed further light on the possible relationship between cybersickness and age. Despite ostensible variation between the symptoms reported by the different age groups, this factor did not produce a significant effect. Those participants who were removed were dispersed across the age groups. While younger participants have traditionally exhibited higher levels of sickness (Paillard et al., 2013), this was not evident in this study. However, there was a difference in the reported enjoyment of the experience. The older age groups rated the experience as significantly less comfortable than their younger counterparts which could be explained by their lack of experience in undertaking hazard awareness tests or due to their limited experience with VR experiences more generally.

4.6 Limitations and future directions

The current study has not focused upon, or reported on, drivers' performance on these tests. While this was not an aim for the current study, it should be a focus of future studies. Hazard tests are often judged on their ability to differentiate between groups of drivers based on a measure of driver risk (e.g. crash history, driving experience, etc.). Follow-on studies must identify whether a test of hazard skill within a VR headset is more effective that traditional presentation modes at differentiating between safe and less-safe drivers. Such studies however will require more stimuli. The current study only included 6 clips per test. If such a test is to be designed to measure participant performance, it is likely to require at least double that number of clips to be presented within a single test. Fortunately, the current results suggest that two tests of six clips (i.e. 12 clips in total) do not significantly increase sickness rates, once participants have passed the initial screening phase.

A further limitation is in the use of the SSQ. While the SSQ effectively identified four participants for removal, preventing any exacerbation of their symptoms through further exposure, it is a relatively lengthy and unwieldy tool to administer on multiple occasions for a single participant. It is likely that shorter versions may be just as effective, and this should also be a focus of future research. Reduction of the SSQ may also allow time for other items to be included. For instance, the demographics form did not enquire after participants' previous VR experience. Including this in future research may disambiguate current findings in the literature (e.g. Weech et al., 2019).

5. Conclusion

This study has provided evidence that VR is a viable format to present hazard awareness tests, with low levels of reported cybersickness compared to a recent meta-analysis of results in this field (Saredakis et al., 2020).

Regarding our concerns that the hazard prediction test would evoke greater sickness, the results suggested the opposite. The hazard prediction test was also preferred by participants, being rated more comfortable and more engaging. Given the additional advantages noted for hazard prediction tests, such as a fairer and more transparent scoring system than that employed in the traditional UK national test (Crundall, et al., 2020; Crundall & Kroll, 2019; Ventsislavova et al., 2019), the current findings support the further exploration of hazard prediction testing in 360-degree environments. While the removal of even one person due to sickness may limit such 360-degree tests in any formal testing situation (such as a national licensing procedure), there are many potential benefits to be gained through hazard awareness assessment and training in less formal settings.

6. References

Agrawal, R., Knodler, M., Fisher, D. L., & Samuel, S. (2018). Virtual reality headset training: can it be used to improve young drivers' latent hazard anticipation and mitigation skills. *Transportation Research Record, 2672*, 20-30.

Aoki, H., Oman, C. M., Buckland, D. A., and Natapoff, A., (2008). Desktop-VR system for preflight navigation training. *Acta Astronautica, 63,* 841-847.

Arns, L. L., & Cerney, M. M. (2005, March). The relationship between age and incidence of cybersickness among immersive environment users. *IEEE Proceedings. VR 2005. Virtual Reality, 2005,* (pp. 267-268). Bonn, Germany: IEEE.

Aykent, B., Yang, Z., Merienne, F., & Kemeny, A. (2014). Simulation sickness comparison between a limited field of view virtual reality head mounted display (oculus) and a medium range field of view static ecological driving simulator (Eco2). Proceedings of the DSC 2014 Europe Driving Simulation Conference (pp.65-71). Paris, France: IFSTTAR.

Benoit M., Guerchouche R., Petit P.-D., Chapoulie E., Manera V., Chaurasia G., Drettakis, G., & Robert, P., (2015). Is it possible to use highly realistic virtual reality in the elderly? A feasibility study with image-based rendering. *Neuropsychiatric Disease and Treatment, 11*, 557–563.

Bessa, M., Melo, M., Narciso, D., Barbosa, L., & Vasconcelos-Raposo, J. (2016). Does 3D 360 video enhance user's VR experience? an evaluation study. *Proceedings of the XVII International Conference on Human Computer Interaction*, (pp. 1-4). New York, NY: Association for Computing Machinery.

Bonato, F., Bubka, A., Palmisano, S., Phillip, D., & Moreno, G. (2008). Vection change exacerbates simulator sickness in virtual environments. *Presence: Teleoperators and Virtual Environments, 17, 283-292.*

Bos, J. E., Bles, W., & Groen, E. L. (2008). A theory on visually induced motion sickness. *Displays, 29,* 47-57.

Brooks, J. O., Goodenough, R. R., Crisler, M. C., Klein, N. D., Alley, R. L., Koon, B. L., et al. (2010). Simulator sickness during driving simulation studies. *Accident Analysis & Prevention, 42,* 788-796.

Cassavaugh, N.D., Domeyer, J.E., & Backs, R.W., (2011). Lessons learned regarding simulator sickness in older adult drivers. In C. Stephanidis (Ed.) *Universal access in human–computer interaction, Context Diversity, vol. 6767* (pp. 263-269). Berlin, Germany: Springer.

Chen, Y. C., Dong, X., Hagstrom, J., & Stoffregen, T. A. (2011). Control of a virtual ambulation influences body movement and motion sickness. *Proceedings of the BIO Web of Conferences, 1,* 1–3.

Classen, S., Bewernitz, M., & Shechtman, O., (2011). Driving simulator sickness: An evidence-based review of the literature. *American Journal of Occupational Therapy, 65,* 179-188.

Crundall, D. (2016). Hazard prediction discriminates between novice and experienced drivers. *Accident Analysis & Prevention, 86,* 47-58.

Crundall, D. E., & Underwood, G. (1998). Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics, 41,* 448-458.

Crundall, D., & Kroll, V. (2018). Prediction and perception of hazards in professional drivers: Does hazard perception skill differ between safe and less-safe fire-appliance drivers? *Accident Analysis & Prevention, 121,* 335-346.

Crundall, D., Chapman, P., France, E., Underwood, G., & Phelps, N. (2005). What attracts attention during police pursuit driving? *Applied Cognitive Psychology, 19,* 409-420.

Crundall, D., van Loon, E., Baguley, T. & Kroll, V. (2021) A novel driving assessment combining hazard perception, hazard prediction and theory questions. *Accident Analysis & Prevention, 149,* 105847.

Davis, S., Nesbitt, K., & Nalivaiko, E. (2014). A systematic review of cybersickness. IE2014: Proceedings of the 2014 Conference on Interactive Entertainment, (pp. 1-9). New York, NY: Association for Computing Machinary.

Duh, H. B. L., Parker, D. E., & Furness, T. A. (2001). An “independent visual background” reduced balance disturbance envoked by visual scene motion: implication for alleviating simulator sickness. *CHI '01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 85-89). New York, NY: Association for Computing Machinery.

Duh, H. B., Parker, D. E., & Furness, T. A. (2004a). An independent visual background reduced simulator sickness in a driving simulator. *Presence: Teleoperators & Virtual Environments, 13,* 578-588.

Duh, H. B., Parker, D. E., Philips, J. O., & Furness, T. A. (2004b). “Conflicting” motion cues to the visual and vestibular self-motion systems around 0.06 hz evoke simulator sickness. *Human Factors, 46*, 142-153.

Fernandes, A. S., & Feiner, S. K. (2016). Combating VR sickness through subtle dynamic field-of-view modification*. 2016 IEEE Symposium on 3D User Interfaces (3DUI)*, (pp. 201-210). Greenville, SC: IEEE.

Freina, L., and Cassana, A., (2015). Immersive vs desktop virtual reality in game based learning. In R. Munvold and L. Kolas (Ed.s) *Proceedings of the 9th European Conference on Games Based Learning* (pp. 195-202). Reading, UK: Academic Conferences and Publishing International Limited.

Forster, Y., Paradies, S., Bee, N., Bülthoff, H., Kemeny, A., & Pretto, P. (2015). The third dimension: Stereoscopic displaying in a fully immersive driving simulator. In Bülthoff et al. (Ed.s) *Proceedings of DSC 2015 Europe Driving Simulation Conference & Exhibition* (pp. 25-32). Tübingen, Germany: Max Planck Institute for Biological Cybernetics.

Gamito, P., Oliveira, J., Morais, D., Baptista, A., Santos, N., Soares, F., et al. (2010). Training presence: the importance of virtual reality experience on the "sense of being there". *Annual Review of Cybertherapy and Telemedicine,* 2010, 128–133.

Gamito, P., Oliveira, J., Santos, P., Morais, D., Saraiva, T., Pombal, M., et al. (2008). Presence, immersion and cybersickness assessment through a test anxiety virtual environment. *Annual Review of Cybertherapy and Telemedicine*, 83–90.

Golding, J. F. (1998). Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain Research Bulletin, 47*, 507-516.

Herrera, F., Bailenson, L., Weisz , E., Ogle, E., and Zaki, J., (2018). Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PLOS One, 13, 10,* e0204494.

Horswill, M. S., & McKenna, F. P. (2004). Drivers' hazard perception ability: Situation awareness on the road. In S.Banbury & S. Tremblay (Eds.), A cognitive approach to situation awareness: Theory and application (pp. 155–175). Aldershot, England: Ashgate.

Horswill, M. S. (2017). Hazard perception tests. In D. L. Fisher et al. (Eds.) *Handbook of teen and novice drivers: Research, practice, policy, and directions*, (pp. 439-450). Boca Raton, FL.: CRC Press.

Horswill, M. S., Hill, A., & Jackson, T. (2020). Scores on a new hazard prediction test are associated with both driver experience and crash involvement. *Transportation Research Part F: Traffic Psychology and Behaviour, 71*, 98-109.

Hosseini, M., & Swaminathan, V. (2016). Adaptive 360 VR video streaming: Divide and conquer. *2016 IEEE International Symposium on Multimedia* (pp.107-110). San Jose, CA: IEEE.

Kennedy, R. S., Drexler, J. M., Compton, D. E., Stanney, K. M., Lanham, D. S., & Harm, D. L. (2003). Configural scoring of simulator sickness, cybersickness, and space adaptation syndrome: Similarities and differences. In L. J. Hettinger & M. W. Haas (Eds.), *Virtual and adaptive environments: Applications, implications, and human performance issues* (p. 247–278). Mahwah, NJ.: Lawrence Erlbaum Associates Publishers.

Kennedy, R. S., Drexler, J., & Kennedy, R. C. (2010). Research in visually induced motion sickness. *Applied Ergonomics, 4*, 494-503.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. The International *Journal of Aviation Psychology, 3*, 203-220.

Keshavarz, B., Ramkhalawansingh, R., Haycock, B., Shahab, S., & Campos, J. L. (2018). Comparing simulator sickness in younger and older adults during simulated driving under different multisensory conditions. *Transportation Research Part F: Traffic Psychology and Behaviour, 54,* 47-62.

Kim, H. K., Park, J., Choi, Y., & Choe, M. (2018). Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics, 69,* 66-73.

Knight, M. M., & Arns, L. L. (2006, July). The relationship among age and other factors on incidence of cybersickness in immersive environment users. In *Proceedings of the 3rd symposium on Applied perception in graphics and visualization* (pp. 162-162). New York, NY: Association for Computing Machinery.

Kolasinski, E.M. (1995) Simulator Sickness in Virtual Environments (Technical Report 1027). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Mabrook, R., & Singer, J. B. (2019). Virtual reality, 360 video, and journalism studies: Conceptual approaches to immersive technologies. *Journalism Studies, 20*, 2096-2112.

MacQuarrie, A., & Steed, A. (2017). Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. *2017 IEEE Virtual Reality (VR)*, (pp. 45-54). Los Angeles, CA: IEEE.

Madigan, R., & Romano, R. (2020). Does the use of a head mounted display increase the success of risk awareness and perception training (RAPT) for drivers? *Applied Ergonomics, 85,* 103076.

Matas, N. A., Nettelbeck, T., & Burns, N. R. (2015). Dropout during a driving simulator study: a survival analysis. *Journal of Safety Research, 55,* 159-169.

Min, B. C., Chung, S. C., Min, Y. K., & Sakamoto, K. (2004). Psychophysiological evaluation of simulator sickness evoked by a graphic simulator. *Applied Ergonomics, 35*, 549-556.

Moran, C., Bennett, J. M., & Prabhakharan, P. (2019). Road user hazard perception tests: A systematic review of current methodologies. *Accident Analysis & Prevention, 129*, 309-333.

Moss, J. D., & Muth, E. R. (2011). Characteristics of head-mounted displays and their effects on simulator sickness. *Human Factors, 53*, 308-319.

Ng, A. K., Chan, L. K., & Lau, H. Y. (2020). A study of cybersickness and sensory conflict theory using a motion-coupled virtual reality system. *Displays, 61*, 101922.

Pai Mangalore, G., Ebadi, Y., Samuel, S., Knodler, M. A., & Fisher, D. L. (2019). The Promise of Virtual Reality Headsets: Can They be Used to Measure Accurately Drivers' Hazard Anticipation Performance? *Transportation Research Record, 2673*, 455-464.

Paillard, A. C., Quarck, G., Paolino, F., Denise, P., Paolino, M., Golding, J. F., & Ghulyan-Bedikian, V. (2013). Motion sickness susceptibility in healthy subjects and vestibular patients: effects of gender, age and trait-anxiety. *Journal of Vestibular Research, 23*, 203-209.

Pope, V. C., Dawes, R., Schweiger, F., & Sheikh, A. (2017). The geometry of storytelling: Theatrical use of space for 360-degree videos and virtual reality. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, (pp. 4468-4478). New York, NY: Association for Computing Machinery.

Pradhan, A.K., & Crundall, D. (2017). Hazard avoidance in young novice drivers: definitions and a framework. In: Fisher, D.L., Caird, J., Horrey, W., Trick, L. (Eds.), Handbook of Teen and Novice Drivers: Research, Practice, Policy, and Directions. Boca Raton, FL: CRC Press.

Prothero, J. D. (1998). The role of rest frames in vection, presence and motion sickness. Doctoral Dissertation, University of Washington. Accessed from http://cumincad.scix.net/data/works/att/diss_prothero.content.04134.pdf (26/01/21).

Prothero, J., & Parker, D. (2003). *A unified approach to presence and motion sickness.* In L. J. Hettinger & M. W. Haas (Eds.), *Virtual and adaptive environments: Applications, implications, and human performance issues* (p. 47–66). Mahway, NJ: Lawrence Erlbaum Associates.

Reason, J. T., & Brand, J. J., (1975). *Motion sickness*. London, England: Academic press.

Rebenitsch, L., & Owen, C. (2016). Review on cybersickness in applications and visual displays. *Virtual Reality, 20*, 101-125.

Rolnick, A., and Lubow, R. E., (1991). Why is the driver rarely motion sick? The role of controllability in motion sickness. *Ergonomics, 34, 7,* 867–879

Ruddle, R. A., & Lessels, S. (2009). The benefits of using a walking interface to navigate virtual environments. *ACM Transactions on Computer-Human Interaction (TOCHI), 16*, 1-18.

Ruddle, R. A., Payne, S. J., & Jones, D. M. (1999). Navigating large-scale virtual environments: What differences occur between helmet-mounted and desk-top displays? *Presence: Teleoperators & Virtual Environments, 8*, 157-168.

Saredakis, D., Szpak, A., Birckhead, B., Keage, H. A., Rizzo, A., & Loetscher, T. (2020). Factors associated with virtual reality sickness in head-mounted displays: A systematic review and meta-analysis. *Frontiers in Human Neuroscience, 14*, 96.

Schmidt, E. A., Kuiper, O. X., Wolter, S., Diels, C., and Bos, J. E., (2020). An international survey on the incidence and modulating factors of carsickness. *Transportation Research Part F., 71,* 76-87.

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI, 3,* 74.

Stanney, K. M., Kennedy, R. S., & Drexler, J. M. (1997). Cybersickness is not simulator sickness. Paper presented at the *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *41*, 1138-1142.

Stanney, K. M., Kennedy, R. S., Drexler, J. M., & Harm, D. L. (1999). Motion sickness and proprioceptive aftereffects following virtual environment exposure. *Applied ergonomics, 30*, 27-38.

Srivastava, P., Rimzhim, A., Vijay, P., Singh, S., and Chandra, S., (2019). Desktop VR is better than non-ambulatory HMD VR for spatial learning. *Frontiers in Robotics and AI, 6,* 50.

Trick, L. M., & Caird, J., (2011). Methodological issues when conducting research on older drivers. In D.L. Fisher et al. (Ed.s), *Handbook of driving simulation for engineering, medicine, and psychology*. Boca Raton, FL: CRC Press.

Ventsislavova, P., & Crundall, D. (2018). The hazard prediction test: A comparison of free-response and multiple-choice formats. *Safety Science, 109*, 246-255.

Weech, S., Kenny, S., & Barnett-Cowan, M. (2019). Presence and cybersickness in virtual reality are negatively related: A review. *Frontiers in Psychology, 10,* 158.

Weidner, F., Hoesch, A., Poeschl, S., & Broll, W. (2017). Comparing VR and non-VR driving simulations: An experimental user study. *2017 IEEE Virtual Reality (VR),* (pp. 281-282), Los Angeles, CA: IEEE.

Wells, P., Tong, S., Sexton, B., Grayson, G., & Jones, E. (2008). Cohort II: A study of learner and new drivers (Road Safety Research Report No. 81). London, England: Department for Transport.

Wetton, M. A., Hill, A., and Horswill, M. S., (2011). The development and validation of a hazard perception test for use in driver licensing. *Accident Analysis and Prevention, 43, 5,* 1759-1770.

Whitney, S. L., Sparto, P. J., Cook, J. R., Redfern, M. S., & Furman, J. M. (2013). Symptoms elicited in persons with vestibular dysfunction while performing gaze movements in optic flow environments. *Journal of Vestibular Research, 23*, 51-60.