RAC Foundation

Mobility • Safety • Economy • Environment

A Comparison of Virtual Reality and Non-Virtual Reality Approaches to Hazard Perception Training & Testing

Does a 360-Degree Environment Provide Tangible Benefits?

David Crundall, Thomas Goodge,
Victoria Kroll, Editha van Loon,
Michael Vernon & Petya Ventsislavova
Nottingham Trent University

August 2021

THE ROAD SAFETY TRUST
Making Roads Safer

Driver & Vehicle Standards Agency

This report has been comissioned by:

The **Royal Automobile Club Foundation for Motoring Ltd** is a transport policy and research organisation which explores the economic, mobility, safety and environmental issues relating to roads and their users. The Foundation publishes independent and authoritative research with which it promotes informed debate and advocates policy in the interest of the responsible motorist.
www.racfoundation.org

The Road Safety Trust (RST) is dedicated to achieving zero deaths and serious injuries on UK roads. To achieve this, the RST provides funding for practical measures, research, dissemination, and education. The RST works with others to use the wealth of knowledge and understanding about what works to keep road safety high on the national and local agenda and influence policy change. The RST shares new knowledge from research and practical interventions across the road safety and wider community to raise awareness and encourage implementation.

www.roadsafetytrust.org.uk

The Driver and Vehicle Standards Agency (DVSA) is an executive agency of the Department for Transport. The DVSA:

- sets the standards for driving and riding and for vehicle safety;
- carries out driving tests;
- approves people to be driving instructors;
- approves MOT testers;
- tests lorries and buses to make sure they are safe to drive; and
- carries out roadside checks on drivers and vehicles and monitors vehicle recalls.

www.gov.uk/government/organisations/driver-and-vehicle-standards-agency

This report has been published by:

RAC Foundation
89–91 Pall Mall
London
SW1Y 5HS

Tel no: 020 7747 3445
www.racfoundation.org

Registered Charity No. 1002705
August 2021 © Copyright Royal Automobile Club Foundation for Motoring Ltd

RAC
Foundation
Mobility • Safety • Economy • Environment

A Comparison of Virtual Reality and Non-Virtual Reality Approaches to Hazard Perception Training & Testing

Does a 360-Degree Environment Provide Tangible Benefits?

David Crundall, Thomas Goodge,
Victoria Kroll, Editha van Loon,
Michael Vernon & Petya Ventsislavova
Nottingham Trent University

August 2021

THE ROAD SAFETY TRUST
Making Roads Safer

Driver & Vehicle
Standards
Agency

About the Authors

Professor David Crundall is a traffic and transport psychologist at Nottingham Trent University. He has published over a hundred academic papers and book chapters in the field, and has conducted research on a wide range of driving safety topics, previously working with the Department for Transport, the Driver and Vehicle Standards Agency, Engineering and Physical Sciences Research Council, Economic and Social Research Council, the Road Safety Trust and many corporate sponsors.

Thomas Goodge is a research assistant at Nottingham Trent University. He has undertaken several projects for sponsors including the development of hazard tests for HGV and bus drivers.

Dr Victoria Kroll is a research fellow at Nottingham Trent University, specialising in the development of hazard perception and prediction tests. She has designed hazard tests for the emergency services and for a recent Department for Transport project. She is also the CEO of Esitu Solutions, a company set up to assess and train commercial drivers in hazard perception skills.

Dr Editha van Loon is a senior research fellow at Nottingham Trent University. She has published over 40 papers on subjects including transport-relevant topics such as autism and driving, eye movements while driving, motorcycle training, and hazard perception. She has a strong interest in programming and developing software and stimuli for transport-based research.

Dr Michael Vernon is a lecturer at Nottingham Trent University, and a cognitive psychologist with an interest in both theoretical and applied domains, including traffic and transport psychology, health and technology, and psycholinguistics.

Dr Petya Ventsislavova is a senior lecturer and a traffic and transport psychologist at Nottingham Trent University. Her interests include cross-cultural hazard perception skills, hazard perception and prediction methodologies, and driving anxiety. She has undertaken driving research in China, Israel, Lithuania, Spain and the UK.

Acknowledgements

The authors would like to thank the RAC Foundation, the Road Safety Trust and the Driver and Vehicle Standards Agency (DVSA) for funding this work. Particular thanks goes to Tanya Qadir, Elizabeth Box, Doreen Lam, Helen Luker and other members of the sponsors' respective teams who have provided invaluable assistance throughout the project. The authors would also like to thank Jellylearn for their assistance in developing stimuli, and our group of approved driving instructors, who provided vital feedback that helped shape the development of our tests.

Disclaimer

This report has been prepared for the RAC Foundation, the Road Safety Trust, and the DVSA by Professor David Crundall and his research team at NTU Psychology, Nottingham Trent University. Any errors or omissions are the authors' sole responsibility. The report content reflects the views of the authors and not necessarily those of the sponsors, and it is the authors who are referred to when "we" and other first-person pronouns are used.

Contents

List of Figures

List of Tables

List of Abbreviations

ANCOVA	analysis of covariance
ANOVA	analysis of variance
AOI	area of interest
CGI	computer-generated imagery
CRIE	comfort, realism, immersion and engagement
DfT	Department for Transport
DoF	degrees of freedom
DVSA	Driver and Vehicle Standards Agency
RAPT	Risk Awareness and Perception Training
SSQ	simulator sickness questionnaire
UVC	ultraviolet C
VR	virtual reality

Foreword

We live in an age where our experiences are increasingly virtual and digital. From online shopping, to remote communication, to hyper-realistic computer games, we can increasingly interact with a lot of what the world – and hence daily life – has to offer from the comfort and privacy of our sofas and desks. Covid appears to have only hastened that trend.

Yet driving remains very much a real-world, hands-on task. And while there is great interest and excitement in, and hopes for, autonomous vehicles, for now when the 40-million-plus of us who have licences take to the road we rely on our human capabilities.

However, could we better use technology to prepare us for the process of driving?

As well as the actual ability to control a vehicle, would-be drivers are also required to pass a theory and a hazard perception test. The latter currently involves a student being shown 14 video clips each of which contains at least one developing hazard that needs to be identified accurately and quickly.

But what if this process could be more immersive? More realistic? More engaging?

That is why the RAC Foundation, the Road Safety Trust and the DVSA commissioned expert researchers at Nottingham Trent University to investigate whether CGI simulations or 360-degree imagery using footage of real roads and traffic could take this aspect of driver training to the next level of realism.

The report which follows describes what the research team found. Its starting premise is that driving is changing – more cars, busier roads, greater reliance on in-car and roadside technology and driver-assist features – and that driver training needs to change with it.

The amendments made to the driving test itself back in 2017 – one of which added a requirement for the learner to follow satnav directions – might seem modest, but in the context of a system that has barely changed for decades they were a welcome addition. From what we have discovered from this work, they will be far from the last changes that we will need to see.

Driving will never be a game – far from it – but if we can use things like gaming technology to improve the learning process then we will all win.

Steve Gooding

Steve Gooding

Director, RAC Foundation

Executive Summary

Hazard perception, past and future

The hazard perception test was developed in the UK, with decades of research demonstrating that it can differentiate between safe and less-safe drivers. The evidence convinced the Government to introduce an official hazard perception test as part of the UK licensing procedure in 2002. The premise is simple: learners watch 14 clips of driving, viewed as if from the driving seat of a moving vehicle. Thirteen clips contain one hazard, while one clip contains two. Hazards might include a pedestrian stepping into the road, or an oncoming car turning across your lane. Viewers must press a button as soon as they see the hazard develop, with faster responses scoring more points. The maximum number of points that can be scored for one hazard is 5, with fewer points awarded for slower responses, and zero points given if the button is pressed too late. Total scores can range between 0 and 75 points, with a score of 44 required to pass the test. In 2015, 1.8 million hazard tests were undertaken, with 85% resulting in a pass mark[1] (compared to 47% of on-road driving tests passed in the same period; see DRT0201[2]). The hazard test is part of the larger theory test, including multiple-choice questions. In 2015, the pass rate for the total theory test was 50%, suggesting that the current UK hazard perception test is the easiest component of the theory test (see DRT5201[2]).

Numerous studies have shown that safer, more-experienced drivers respond faster to such hazards than do inexperienced or unsafe drivers. One potential problem with the methodology employed by the UK hazard perception test, however, is that clips are usually presented on a computer monitor providing between 40 and 60 degrees of visual angle. This may exaggerate the hazard perception skills of poor drivers by focusing their attention on areas where hazards will appear (primarily, the road ahead). It also removes the opportunity to display hazards that originate from outside this forward cone of vision (e.g. an undertaking vehicle).

Virtual reality (VR) headsets may, however, provide an excellent opportunity to address these issues and improve the assessment and training of drivers' hazard perception skills. Presenting 360-degree hazard clips in a VR headset provides viewers with the opportunity to look wherever they want, such as into side roads as they pass them, and to check side mirrors and blind spots for other road users. Few studies have assessed the predicted benefits of 360-degree presentation for hazard perception assessment and training. The aim of the current project was to provide this evidence by creating and comparing hazard tests and training materials, presented in VR headsets and on single screens.

[1] https://www.whatdotheyknow.com/request/statistics_of_people_that_obtain

[2] https://www.gov.uk/government/statistical-data-sets/driving-test-statistics-drt#driving-test-and-motorcycle-test-pass-rates-drt01

Study 1: Will VR hazard tests make people sick?

A small number of VR users typically report nausea, dizziness, and other symptoms, collectively termed ‘cybersickness’. The first study was intended to assess the level of sickness that might be evoked by a hazard perception test presented in 360 degrees via a VR headset.

First, we recorded footage from a moving vehicle using a 360-degree camera. This footage resulted in the selection of 12 clips, each containing a naturally occurring hazard. These clips were edited to create two variants of a hazard test: the hazard perception method employed by the UK Driver and Vehicle Standards Agency (DVSA), and a modified version termed the ‘hazard prediction test’ (see Box 1). The hazard prediction test offers advantages over the traditional test format, but because it involves each clip stopping abruptly during perceived motion, we were concerned that the prediction test might evoke greater sickness than the more typical hazard perception test.

We recruited 77 participants to undertake the tests using a VR headset and measured their cybersickness symptoms with the simulator sickness questionnaire. Four participants were removed owing to excessive cybersickness symptoms during a practice session (5%), a proportion which is low in comparison with the literature.

Contrary to our concerns, the remaining participants rated their symptoms as lower in the hazard prediction test than in the hazard perception test. Participants also rated the hazard prediction test as more comfortable and more engaging than the hazard perception test. Older drivers felt less comfortable overall inside both tests, but there was no clear influence of participants’ age on cybersickness symptoms.

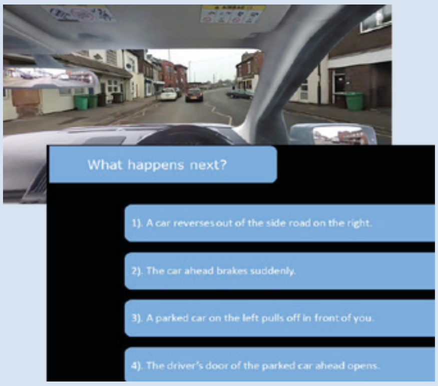
Box 1: Hazard perception vs hazard prediction

The hazard perception test relies on ‘scoring windows’. These are segments of clips where a hazard is judged to be developing, and therefore a response in this time window is considered a ‘hit’, and scores points according to how quickly the response was made once the window opened.

Using scoring windows can cause problems, however. It is possible for excellent drivers to anticipate a hazard and press slightly too soon, scoring zero points. The opposite problem may also occur: experienced drivers may spot a hazard early but wait to respond, because the danger is not yet sufficient to challenge their driving skills. Both errors can unfairly penalise safe drivers.

Our favoured alternative is the hazard prediction test. Drivers see the clip up to the point when the hazard begins to develop; then the clip suddenly stops, and the scene is entirely hidden from the driver’s view. We then ask “What happens next?” Viewers choose an answer from four options.

Direct comparisons of hazard perception and hazard prediction methodologies have previously suggested that the latter is better at differentiating between safe and less-safe drivers (Crundall & Kroll, 2018, Ventsislavova et al., 2019).



What happens next?

1). A car reverses out of the side road on the right.

2). The car ahead brakes suddenly.

3). A parked car on the left pulls off in front of you.

4). The driver’s door of the parked car ahead opens.

Fig (i): Hazard prediction clips play until the hazard begins to develop; the clip then stops are participants are asked “What happens next?”

The results demonstrated that it was possible to create a VR test that evoked cybersickness in only a small number of viewers. Furthermore, lower sickness scores for the hazard prediction test allowed us to use this test format for the subsequent studies.

Study 2: Is a VR hazard test better than the same test shown on a 2D screen?

Fresh footage was recorded from a moving vehicle to create 24 new hazard prediction clips that could be presented either in VR or on a 2D screen (see Box 2). Sixty-seven participants, a mixture of experienced and novice drivers, were recruited to view the clips. Given novices' higher crash risk, we expected these inexperienced drivers to perform worse on a hazard prediction test than the experienced group. The important question, however, was whether the gap between novice and experienced driver performance on the hazard prediction test was greater in the VR version or the 2D single-screen version.

Box 2: Creating VR hazard prediction tests

For Study 2, we filmed footage from a moving vehicle using a 360-degree camera. Additional cameras captured rearward views, which were then synchronised with the 360-degree footage and edited into the mirrors of a graphic overlay of a car interior. For Study 3, we created similar clips using computer-generated imagery (CGI).

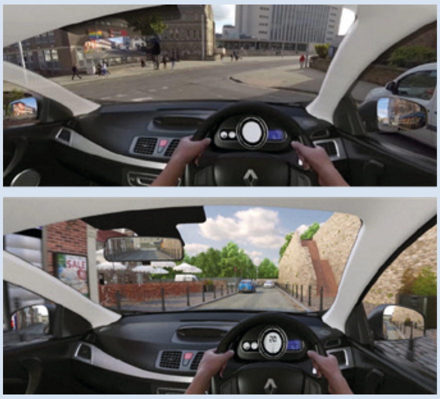


Fig (ii): Screenshot of a VR hazard prediction clip created from recorded footage (top panel) and a screenshot of a similar clip created using CGI (bottom panel)

Two participants were removed from the study owing to sickness (3% of the total). After viewing both tests, participants rated the 360-degree test more highly on dimensions of realism, immersion, and task engagement. Both tests contributed to the finding that experienced drivers predicted more hazards than novices did, though there was slight evidence that the 360-degree test might have been more effective at differentiating between these groups (see Box 3).

Study 3: Let's try that again, but with CGI

Most research groups record video footage to create hazard tests. In 2015, however, the DVSA replaced their video clips with clips consisting of computer-generated imagery (CGI). This change was undertaken to reduce the production costs of the clips, and to improve their longevity (on the assumption that it is cheaper to edit existing CGI clips to update car models etc. than it is to undertake new filming). Programmed hazards also allow a level of control not possible with hazards captured in video footage. To ensure the relevance of our research to the DVSA, we replicated Study 2, but using ten CGI clips designed by our researchers and programmed by the same company who produces the official DVSA clips (see Box 2).

There were 125 participants recruited for the study. They were split into four groups: novices assigned to either the 360-degree CGI test or the single-screen CGI test, and experienced drivers who were also divided between the two tests.

The rationale remained the same as Study 2: to discover whether the VR test would be better or worse at differentiating the experienced from the novice driver groups than the 2D single-screen test.

Eight participants were removed owing to cybersickness (6%). The remaining participants produced results that mirrored Study 2's results closely. Experienced drivers outperformed novices in predicting hazards, with both tests showing evidence of this effect. Again, there was slight evidence that the 360-degree test was better at differentiating between the driver groups than the single-screen test (see Box 3). Following their test, participants were shown the clips in the alternate presentation mode and their preference ratings for VR and the 2D single-screen test were compared. Participants rated the VR test more highly for realism, immersion and engagement. They did, however, also report that the VR test was less comfortable than the single-screen version, though the size of this negative effect was small.

Box 3: Comparing VR and 2D hazard prediction tests

For both Study 2 and Study 3 we compared both novice and experienced drivers' performance on 360-degree clips presented in VR to performance on the same clips presented on a 2D single screen. Overall, experienced drivers outperformed the novices as expected, and there was some evidence to suggest that the VR tests were better at differentiating between these two groups.

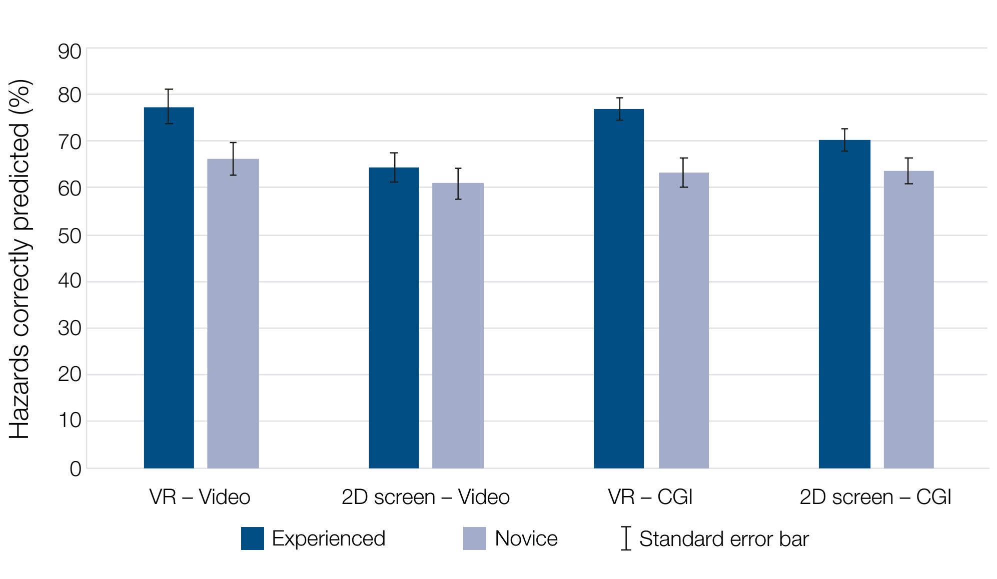


Fig (iii): Data from Study 2 and 3 that demonstrates the gap between experienced drivers' performance and that of the novices

Study 4: Is VR better for training participants?

This study aimed to identify whether there was any benefit to providing training materials in a VR headset compared to doing so on a single screen. Training materials were developed from the CGI hazard clips used in Study 3. Expert voiceovers, additional warning graphics, and a virtual satnav display (to provide a plan view of the hazard) were some of the techniques applied to the clips to help drivers understand where they should look and why (see Box 4). These training clips were presented to one group of drivers in a VR headset, and to another group of drivers on a 2D single screen. A third no-training group was recruited as a control. The impact of the training was assessed from participants' performance on our video-based hazard prediction test presented in a VR headset (adapted from Study 2), and on how well they navigated a virtual route containing ten hazards in a driving simulator.

Ninety-nine participants were recruited and split into the three training groups. Three people were removed owing to sickness symptoms (3%), though two of these participants developed their symptoms on the driving simulator. Only one participant was removed as a result of cybersickness within the VR headset (1%).

Analyses showed that while VR-trained drivers had the highest post-training hazard prediction score, and the control group had the lowest score on the same test, this effect did not reach the threshold of significance. When drilling down into the training effect at an individual clip level, it was clear that training improved prediction accuracy on some clips but not others. When those clips that showed evidence of training benefit (i.e. improved scores) were compared with the training material, there were clear similarities between what was seen during training on the CGI clips and the hazards in the subsequent video-based test. This is an example of 'near transfer' of training, where training improves performance in situations that are highly similar to the training environment. Unfortunately, evidence of extrapolation of training benefits to dissimilar scenarios ('far transfer') was not found.

The comparison of the performance of the training groups on the driving simulator revealed that both training conditions (VR headset and 2D single screen) resulted in a reduction in drivers' lateral variation (e.g. weaving, swerving or drifting). The VR training was most effective at reducing steering wheel error in simulator driving (which is closely linked to lateral variation). This possibly reflects drivers' improved ability to anticipate the need to change lateral position (e.g. spotting a hazard earlier allows a driver to make a smaller adjustment to the lateral position, while still avoiding the danger). There was also weak evidence that VR-trained drivers chose to drive at slower speeds, with the possible suggestion that hazard training had made our drivers more risk-averse.

Study 4 suggested that there is potential benefit for training drivers in hazard perception using a VR headset, but that the CGI training clips alone were not sufficient to create a robust effect. It was recommended that future training efforts should employ multiple instances of specific hazards, starting with less-complex CGI scenarios, before moving to more-complex video-based hazards. Once drivers have gained basic understanding from the CGI clips, and have subsequently been coached to apply this to richer video-based scenarios, they should subsequently reap the benefit of this training in real-world environments.

Box 4: Training hazard perception

Training clips were developed from the CGI clips used in Study 3. They included an expert voice-over telling them where they should have looked and why, supported by additional markers (e.g. highlighting important areas of the scene). A virtual satnav was also included to allow a top-down view of the hazardous scenario. These training clips were presented to participants either in a VR headset or on a 2D computer screen. A control condition involved another group of participants who received no training at all. Training benefits were subsequently measured in terms of participants' performance on a video-based hazard prediction test in a VR headset (adapted from Study 2) and how well they navigated a virtual route in a driving simulator that contained ten hazards.

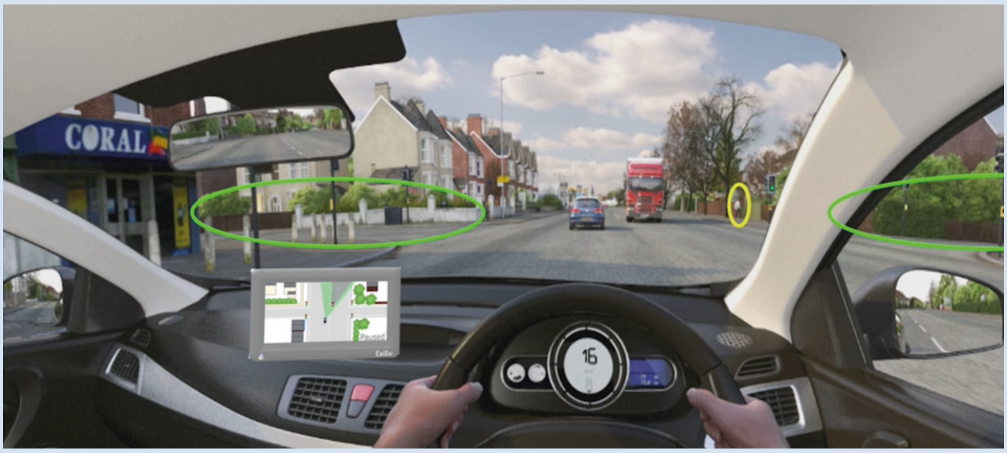


Fig (iv): A screenshot from a CGI training video adapted from one the clips used in Study 3

Study 5: Do participants prefer CGI clips or naturally recorded video clips in their VR test?

The two main hazard prediction tests created for this project (Study 2 using video clips and Study 3 using CGI clips) gave remarkably similar patterns of results. For future research and application, however, we wanted to know whether a new cohort of participants preferred the video-based test or the CGI-based test. This study directly compared participants' views of the CGI and video-based tests when presented in a VR headset (this study used only VR tests and did not include a 2D single-screen condition). Thirty-four participants undertook both tests comprising ten CGI clips and ten video-based clips. None of the participants were removed from the study owing to sickness symptoms, though two were lost as a result of equipment failure.

Participants reported the video-based test to be more realistic, and to have greater clarity and visual complexity. When asked which test they thought would be better for assessing their hazard perception skills, most participants (56%) chose the video-based test over the CGI test.

Fifty-nine percent of participants thought that the video-based test offered the best assessment of their ability to predict hazards ,whereas 22% thought the, CGI test to be better, as those tests currently stood. Nineteen percent were ambivalent. When asked to imagine how much better both tests might get in the future, the preference for video increased (to 63% vs 22%; see Box 5).

Box 5: Comparing the VR tests

Participants were asked to rate the tests on a seven-point scale to reflect their preference for the video-based test (the low end of the scale) or the CGI test (the high end of the scale). There was a clear preference for the video-based test, both in its current state, and when considering the potential for future improvements to both tests.

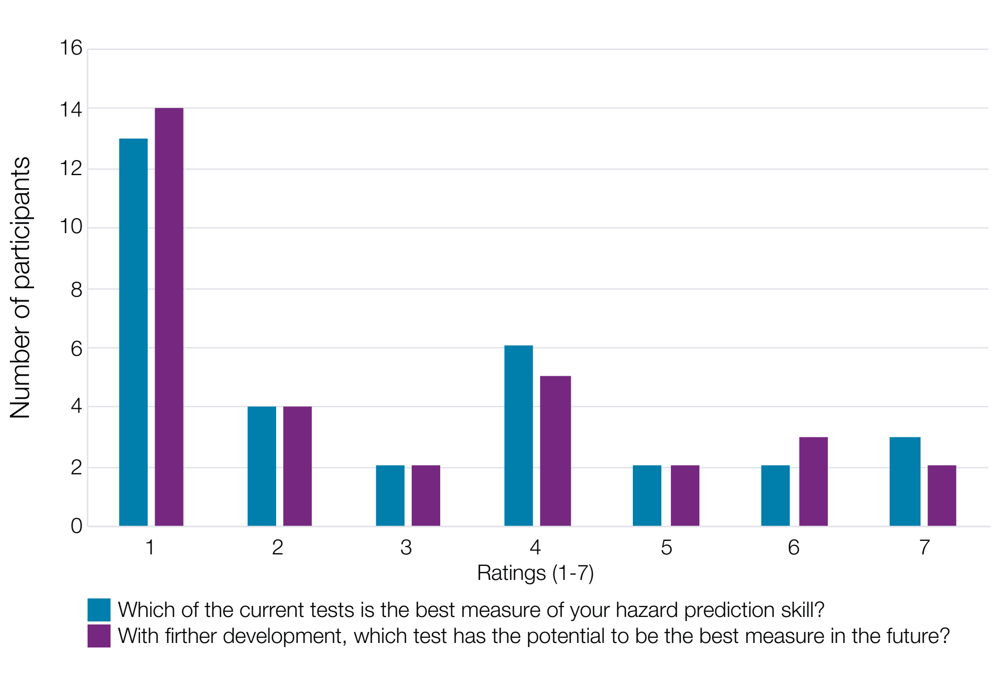


Fig (v): Participants' ratings reflecting their preference for either the video-based or the CGI hazard prediction test

Study 6: Assessing preferences through an Oculus Go Store app

Study 6 was not originally planned but came about as a response to the university putting a halt to research during the COVID-19 pandemic in 2020. To collect data without face-to-face contact, we developed an app that was released in the Oculus Go Store. The 'Hazard Perception VR' app was designed as an online version of Study 5, containing all the same clips and questions, and recording all the same data as the laboratory equivalent. Fortunately, in the event, we were able to complete Study 5 in the laboratory, which actually then provided an unexpected opportunity to compare the views of participants tested in the lab with those held by a sample of VR hobbyists using the app – who might be expected to be more exacting than the average VR user in their expectations of what such an app should offer.

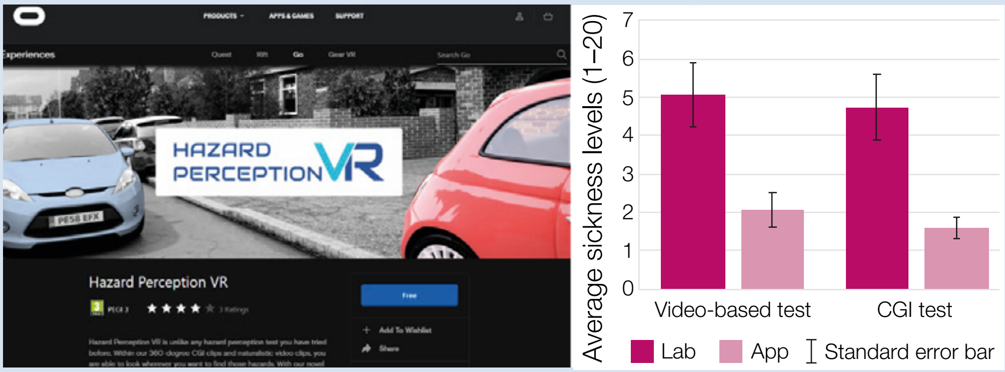
Launched on 12 November 2020, the app was downloaded over 350 times in the first few months, though many users did not go on to complete the whole study. It is understandable that, confronted with an online consent form, a demographic questionnaire and copious details about their rights as participants (all required by the university when collecting data), the majority of these casual users decided that the barriers to using the app were off-puttingly high. As a result, only 20 people completed the full test.

Comparing app users' data with data collected in the lab, we found app users to report lower sickness symptoms and higher comfort ratings, possibly reflecting a self-selection effect (people are unlikely to buy – or continue to use – a VR headset if they are afflicted by cybersickness). Alternatively, regular VR users may have become acclimatised to the immersive environments. App users were also more forgiving of the CGI clips' limited realism, though they still rated the CGI clips as having lower clarity and complexity than video.

Despite the high dropout rate, the initial number of downloads is promising. The intention is to update the app and release a new version for the latest (and more popular) Oculus Quest headset, which should result in a wider uptake.

Box 6: Measuring the behaviour of VR users outside the laboratory

Study 6 was a replication of Study 5 with data collected through our 'Hazard Perception VR' app, which was made freely available through the Oculus Go Store in November 2020. These native VR users also preferred the video-based test. Though sickness ratings were low in both Study 5 and Study 6, the native VR users also reported much lower ratings for cybersickness.



Hazard Perception VR is unlike any hazard perception test you have tried before. Within our 360-degree CGI clips and naturalistic video clips, you are able to look wherever you want to find those hazards. With our novel

Fig (vi): Landing page in the Oculus Go Store for the 'Hazard Perception VR' app (left panel); lower sickness ratings of native VR users using the app (Study 6) than those of the more general sample of participants recruited for the lab study (Study 5) (right panel)

Conclusions

These studies have demonstrated that it is feasible to create effective 360-degree hazard tests for presentation within VR headsets which can differentiate between less-safe novice drivers and safer, more-experienced drivers. No other studies have compared VR hazard tests with *identical* tests presented on a 2D single-screen monitor, and we believe that this is the first study to demonstrate these benefits within this domain.

Specifically, we have reported that:

- **It is possible to design hazard tests presented in VR which evoke low levels of cybersickness symptoms**, and we have offered explanations for why this might be the case. While a single case of cybersickness might prevent a VR test being used at a national level because of equality-of-access issues, the low ratings raise the possibility of using VR to identify the training needs of drivers in less-formal contexts.
- **Our hazard prediction test provoked fewer cybersickness symptoms than our hazard perception test**, suggesting that this could be a more appropriate methodology for a VR test.
- **Participants who are most susceptible to cybersickness can be screened out at an early stage** – in our case by using a two-minute practice clip.
- **Both video and CGI clips can create hazard prediction tests that successfully differentiate between novice drivers and more-experienced drivers.** The literature considers this to be a validation criterion for such tests. **The evidence also suggests that in VR tests, the gap in performance between novice and experienced drivers may be greater than in 2D tests.**
- Clear training effects are harder to demonstrate. **Improvements on subsequent hazard prediction performance appear to be limited to those assessment scenarios that are very similar to the training scenarios.** We have recommended an iteration to future training efforts that will build on the evidence here and, hopefully, improve future training benefits.
- **Our participants preferred the video-based clips.** This is possibly due to the greater levels of complexity and realism inherent in the videos. Recommendations are made for the future use of video and CGI clips in hazard assessment and training.
- **Useful data can be collected via the Oculus Go Store, and the results from such native VR users mirror those collected in the laboratory** in several instances. One notable exception is that **owners of VR headsets report much lower levels of cybersickness** compared to participants who are recruited for laboratory research (who are less likely to own their own headset).
- Finally, one of the strongest effects revealed by these studies is that **participants have clear preferences for the 360-degree tests over single-screen versions**. Their enthusiasm for VR assessment and training, in terms of perceived realism, immersion and engagement (but not comfort) is, arguably, reason enough to pursue this route to improved driver safety. Such increased levels of engagement may even encourage some drivers, who might not have previously considered it, to undertake voluntary training in the privacy of their own VR headset.

1. Introduction

Head-mounted virtual reality (VR) has seen a step change in quality and application over recent years (Slater & Sanchez-Vives, 2016). This has led to great excitement among a range of organisations with responsibilities for transport safety. Notable VR-based attempts at improving road safety include:

- AT&T's "It can wait" campaign to reduce mobile phone distraction (a 360-degree video of driving, including hazards, that plays while the driver interacts with a mobile phone);[3]
- Road Safety Scotland's 'Don't Risk It' VR video (a 360-degree video of a test drive that culminates in an unexpected hazard);[4] and
- Leicester Fire & Rescue Service's 'Virtual Fatal 4 360' video of a car crash and its aftermath.[5]

Interest is also growing in the use of head-mounted VR as a training tool, with delivery company UPS being an early adopter.[6] These VR tools theoretically offer huge potential for safety interventions, especially in the field of hazard perception.

[3] https://www.youtube.com/watch?v=xKTLXU-tE5U
[4] https://www.youtube.com/watch?v=hnWgEGVjlak
[5] www.leics-fire.gov.uk/your-safety/road-safety/vf4-360/
[6] www.youtube.com/watch?v=fypGVcmWpnU

Hazard perception refers to the skill of identifying an on-road hazard in sufficient time to avoid a collision. The traditional hazard perception 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 press sooner than less-safe drivers, and the evidence is so consistent that researchers argue this to be the most clear-cut cognitive skill relating to driver safety (e.g. Horswill, 2017). Decades of research led the UK Department for Transport (DfT) to introduce a hazard perception test as part of the licensing procedure in 2002, with subsequent evidence suggesting that this has reduced collisions (Wells et al., 2008).

Hazard perception tests have not been without their critics, however. One criticism is that drivers are presented with a vastly restricted view of the world outside the vehicle (without even the opportunity to look in their cars' mirrors). We know that wider fields of view and mirror information can change the way in which drivers respond in such tests (Shahar et al., 2010), meaning that presenting hazards in a 360-degree environment could evoke more ecologically valid driver behaviour – that is, behaviour that reflects real-world demands. But would 360-degree hazard perception tests, akin to the AT&T and Road Safety Scotland examples, be worth the additional cost and effort?

Unfortunately, while there are hundreds of research papers detailing a broad range of VR applications, relatively few assess the benefits that VR provides over and above more conventional presentation modes (e.g. presenting video content on a single screen). Of the few exceptions that have been reported, Ruddle et al., (1999) compared participant performance on a navigation task via a desktop system with the same via a head-mounted VR. They found no overall task performance differences, though VR users were better at estimating distances between waypoints. This was possibly due to participants' increased likelihood of stopping during the VR navigation task and inspecting non-relevant aspects of the scene (see also Ruddle & Lessels, 2009). More recently, MacQuarrie and Steed (2017) presented participants with video clips (e.g. horror, documentary, etc.) in both VR headsets and on a single screen, with the former leading to reports of greater enjoyment and improved spatial awareness.

Regarding driving in VR, there has been a small but promising flurry of recent research, though the evidence it presents in favour of the VR modality is mixed. For instance, Aykent et al. (2014), Forster et al. (2015) and Weidner et al. (2017) have all published studies comparing driving behaviour in VR headsets with more traditional presentation modes. They did not find driving behaviour to change with the more immersive VR presentation, but all three studies noted increases in simulator sickness (or 'cybersickness'), often defined by a range of symptoms including dizziness, nausea and increased sweating, caused by VR immersion.

In a recent comparison of a fixed-base driving simulator with a VR-based driving simulator, Mangalore et al. (2019) found that drivers were equally likely to spot certain types of hazard in either platform, though as their fixed-base simulator provided a 330-degree wraparound view, the VR headset was unlikely to have provided a viewing advantage.

Indeed, the reported field of view of their HTC Vive (110 degrees under optimal conditions) is less than the unencumbered horizontal field of view of a human observer and could therefore be said to be more restrictive than the viewing position within the fixed-base simulator. Nonetheless, both platforms found experienced drivers to spot more of these hazards than a novice driver group. Interestingly, they observed no significant difference between sickness scores across the two platforms. While overall rates of sickness were relatively low (<10%), their middle-aged experienced drivers reported higher sickness scores on both platforms than the younger drivers, which is consistent with previous research and suggests a relationship between age and sickness (Brooks et al., 2010; Keshavarz et al., 2018).

Finally, two recent studies have attempted to deploy Risk Awareness and Perception Training (RAPT; e.g. Fisher et al., 2004) in VR headsets. In a UK study, Madigan and Romano (2020) compared three types of RAPT training using still images on a single screen, still images in a VR headset, and dynamic images in the VR headset. They found the last of these three training conditions produced the greatest training benefit, as measured by a version of the hazard perception test provided by the Driver and Vehicle Standards Agency (DVSA). Despite the promising results for the VR-training modality, the effect was confounded by the increased exposure time that drivers had in the dynamic VR condition. Furthermore, there was no direct comparison of VR and single-screen conditions using dynamic footage. The effect in favour of VR training was mirrored in a US study that compared VR and single-screen RAPT training which found greater training benefits in the VR condition (Agrawal et al., 2018). Unfortunately, the content between the VR and single-screen training interventions differed, thus it is difficult to conclude that the superiority of the VR condition was due to the presentation mode alone. This latter study also demonstrated low levels of sickness in the VR headsets, with only one participant (out of 36) being removed owing to excessive symptoms.

While VR promises much in terms of immersion, and the evocation of more naturalistic behaviour, the evidence in terms of a benefit for driving safety interventions is relatively mixed, and the research field is missing direct comparisons between hazard awareness in VR and single-screen modalities, for both assessment and training purposes. The current project aimed to plug this evidential gap, by developing several 360-degree hazard tests (and associated training materials) and comparing them with single-screen equivalents. If the benefits of 360-degree presentation can be clearly documented and contrasted against the potential costs of cybersickness, we will be able to judge the impact of such investment. This should ensure that future driving safety interventions are more cost-effective.

By means of a series of five studies, we aimed to address the following questions:

Study 1: What levels of cybersickness symptoms are generated by different types of hazard perception test when presented in VR? This study was considered important to help decide upon the type of hazard test to be used in the subsequent studies. It involved immersing participants in two (slightly different) tests, to assess whether one was more likely to evoke sickness symptoms than the other.

Study 2: Is a 360-degree hazard test, comprising naturally recorded video, better able to differentiate between safe and less-safe drivers than a single-screen version? Hazard tests are often validated by assessing whether they can detect a difference between driver groups who are known to vary on a safety-related measure, such as driving experience or crash history. This study aimed to identify whether the VR variant was more effective at differentiating between such driver groups than a non-VR version.

Study 3: Is a 360-degree hazard test, comprising computer-generated imagery (CGI), better able to differentiate between safe and less-safe drivers than a single-screen version? This study was identical to Study 2, except that it used CGI rather than natural video. While less realistic and complex than natural video, CGI provides more control over the stimuli, and a less-juddery viewing experience than naturalistic video. If Study 2 failed to find a VR benefit, it was possible that the CGI clips in Study 3 would.

Study 4: Is a VR-training environment more effective than a single-screen one at improving drivers' hazard perception skills? This study compared training in VR and single-screen environments, with a no-training control condition, using a hazard test and a simulated drive as outcome measures.

Study 5: Do participants prefer viewing CGI or natural video when engaging with a hazard test in a VR headset? If either mode (CGI or video) shows a benefit in earlier studies, will participants accept this type of test? Will they prefer the more realistic video, or the smoother and less-cluttered experience of the CGI clips?

2. Study 1: A Comparison of Cybersickness Symptoms Across Two Variants of a Hazard Test

2.1 Introduction

2.1.1 Different types of hazard test

The first step in designing a hazard test is to decide which of the many measures of hazard skill should be used. A recent review of nearly 50 research publications in the field of hazard perception (Moran et al., 2019) identified over 100 different measures used to assess hazard perception skill. These tests also differ in many other ways: some use naturalistic hazards, while others record staged events; some create hazards using CGI instead of relying on video footage; some tests can be as simple as presenting static images to drivers, while others might use highly immersive driving simulators (Crundall et al., 2021).

The base-level hazard perception methodology is perhaps best encapsulated in the official UK test. This presents viewers with clips filmed as if from the driver's perspective (once consisting of video recordings of staged events but updated to CGI clips in 2015). Participants watch the clips and press a button as soon as they spot a developing hazard. If the response falls within a predetermined temporal window, it is awarded points. The scoring window is segmented into five periods of equal length. Responses that fall in the first segment score five points. Responses that fall in the second segment score four points, and so on until the scoring window closes. Responses that fall before or after the scoring window score zero points. Arguably, any study attempting to assess the potential for VR hazard perception testing should consider emulating this traditional push-button approach.

There are, however, several potential pitfalls with using simple response times to calculate a score. For instance, a test's validity relies on the positioning of the scoring windows. If it is possible for exceptionally good drivers to anticipate and respond to the correct hazard before the scoring window opens, then the test will unfairly penalise the safest drivers (Crundall, 2016; Pradhan & Crundall, 2017). The opposite problem may also be noted: when measuring the performance of highly competent drivers, these participants may notice a hazard early, but hold off responding because the hazard poses no threat to their superior skills at this point. Only when the hazard gets closer might these highly trained drivers consider that the danger is now worth acknowledging. This was found with pursuit-trained police drivers: physiological measures indicated that they were aware of hazards sooner than a control group, yet their explicit responses were no faster (Crundall et al., 2003).

A further problem with the traditional hazard perception approach is that it is essentially reactive. While instructions might suggest that drivers should anticipate developing hazards, the yardstick by which they are measured is the speed of response once something has happened. This implicitly encourages late reactive responses to hazards, rather than proactive anticipation.

One alternative to this is the *hazard prediction test* (also known as the 'What happens next?' test; Jackson et al., 2009). Instead of measuring response times to hazards, the test records drivers' ability to anticipate imminent hazards: each clip is suddenly occluded at hazard onset, and participants are asked "What happens next?" Drivers must then choose between four text options presented on screen (Ventsislavova & Crundall, 2018). In direct comparisons of hazard perception and hazard prediction methodologies, the latter has proved more effective when comparing novice and experienced drivers across several countries (Ventsislavova et al., 2019), and in distinguishing high-risk from low-risk professional drivers (Crundall & Kroll, 2018). A recent study by Horswill, Hill and Taylor (2020) also linked better hazard prediction performance with fewer self-reported crashes in typical drivers.

While other versions of the hazard perception test exist, we have previously rejected several of them on theoretical and practical grounds (Crundall, 2016), and argue that hazard prediction offers a viable alternative (or supplement) to hazard perception testing (Crundall et al., 2021).

There is, however, a potential problem with translating the hazard prediction test into a 360-degree test to be presented in VR. The use of sudden changes in scene (such as the abrupt occlusions used in the hazard prediction test) are not recommended in VR presentations, as they may exacerbate the symptoms of cybersickness (Bonato et al., 2008).

2.1.2 Measuring cybersickness

Researchers involved in driving stimulator research have long acknowledged the existence of simulator sickness (Kolasinski, 1995; Brooks et al., 2010), with sensory conflict often raised as one possible explanation for such effects (Bos, Bles, and Groen, 2008). For instance, one may perceive the visual input associated with turning a corner in a simulator, but the vestibular system (the apparatus of the inner ear involved in balance) will not register any associated physical movement.

Some of the typical symptoms associated with simulators, including disorientation, nausea, and blurred vision, are exacerbated in VR headsets (Kennedy et al., 2003; Kim et al., 2018). The term ‘cybersickness’ was coined to refer specifically to such symptoms evoked through VR immersion. Regardless of the distinction between cybersickness and general simulator sickness, both are typically measured on the same scale (Rebenitsch & Owen, 2016), with the simulator sickness questionnaire (SSQ) being the most popular measure (Kennedy et al., 1993). The SSQ is a 16-item questionnaire that measures the severity of symptoms such as nausea, sweating and fatigue on a four-point scale (from “none” to “severe”). The SSQ is typically administered to participants at various points during immersion to ascertain whether any symptoms are worsening. This provides an indication as to whether a participant should be removed from a study. For the current study, the SSQ score also acts as the primary dependent variable in our attempt to ascertain which of the two hazard test variants produce the greatest cybersickness symptoms.

As noted above, the hazard prediction test offers advantages over the traditional hazard perception approach, but it may not fare well when presented via a VR headset. Study 1 was therefore designed to assess whether an initial hazard prediction test presented in VR (including abrupt occlusions) is likely to evoke greater sickness symptoms than the traditional response-time approach used by the official UK hazard perception test. By comparing the sickness symptoms evoked by both test variants (as measured by the SSQ), we can determine whether to proceed with the use of hazard prediction in the design of our subsequent 360-degree hazard tests, or whether later studies should rely on a more traditional hazard perception approach.

2.2 Method

2.2.1 Participants

The first study recruited 77 drivers, split across different age groups (17–25, 26–35, 36–45, 46–55, 56+), though four participants were withdrawn from the study owing to high reported sickness levels, and one further participant was removed as a result of data loss (owing to equipment failure). The demographic details of the participants in each age group, and those participants who were removed owing to sickness, are given in Table 2.1.

Table 2.1: Demographics of all participants who completed Study 1

Group	N	Gender	Mean Age (years)	Mean Driving experience (years since passing driving test)
17–25	21	18 females	20.3	2.6
26–35	12	6 females	30.4	10.3
36–45	13	9 females	39.8	15.9
46–55	9	5 females	50.6	29.8
56+	17	6 females	69.8	49.5

Source: Authors' own (Study 1)

2.2.2 Design

A 2 × 5 × 5 mixed design was employed, with five age groups of drivers undergoing both a hazard perception and a hazard prediction test. The third independent variable was the point in time during the study that the drivers' sickness symptoms were assessed. SSQ scores were collected at five points throughout the study: baseline; after a brief acclimatisation in the VR headset (40 seconds), after a longer acclimatisation in the headset (at 2 minutes, 13 seconds), and after both the hazard perception and hazard prediction tests (which were presented in a counterbalanced order).

Each test contained six clips that were selected for inclusion if they contained a clear hazard. Half of the participants in each age group saw Hazards 1–6 (see Table 2.2) in the hazard perception format, while Hazards 7–12 were presented as hazard prediction clips. The other half of the participants saw Hazards 1–6 in the prediction test and Hazards 7–12 in the perception test. Ideally, clips would have been presented in random order within their tests, but the software used to control the experiment (Tobii Pro Lab) did not offer that function at this time. Instead, two orders of clips for each test (sequential and reversed) were factored into the counterbalancing schedule (ensuring that all conditions were seen an equal number of times in their various permutations).

The primary dependent variable for this study was the level of cybersickness symptoms as recorded by the SSQ. Participants were also asked to rate the two tests for comfort, realism, immersion and engagement (the 'CRIE' questions) on ten-point scales. Though not a primary focus of the study, behavioural responses to the two tests were also collected, and eye movements were measured via an eye tracker built into the VR headset.

2.2.3 Stimuli

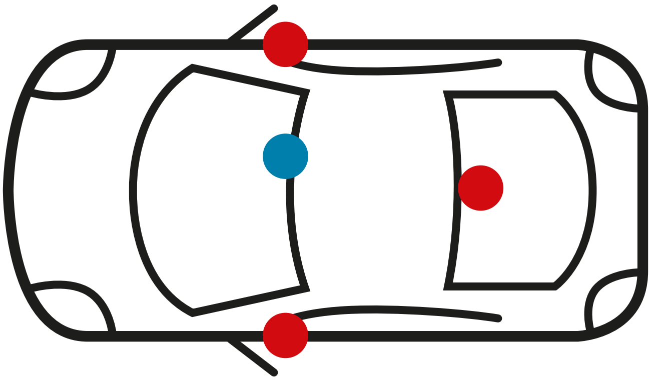
2.2.3.1 Creating the hazard perception and hazard prediction tests

Video footage of real driving was recorded from a Garmin VIRB 360 action camera mounted on a Vauxhall Corsa. This was placed at the top of the windscreen, directly above the driver, to better reflect the driver's perspective. For mirror views, three GoPro HERO4 cameras (1,080p, 16:9 ratio, wide-angle setting) were mounted externally using suction mounts aligned with the mirrors but positioned to avoid obstruction for the driver. Two of these cameras were mounted on the doors to capture side-mirror views. One further camera was positioned on the rear of the vehicle to capture the scene that would normally appear in the rear-view mirror. All cameras were tethered to the vehicle for safety. The locations of these cameras on the film car can be viewed in Figure 2.1. Footage was collected over a two-month period from November 2018 to December 2018 across Nottinghamshire and Nottingham city centre.

A team of traffic psychologists reviewed the footage and selected 12 clips (see Table 2.2), based on the following five principles:

- **a.** the clip contains a hazard of sufficient danger to warrant a change in driver behaviour (i.e. in speed or in lane position) to reduce the possibility of a crash;
- **b.** the hazard can be predicted from a precursor (i.e. a clue to the upcoming hazard, e.g. a pedestrian walking towards a crossing);
- **c.** the hazard had a clearly defined onset (e.g. the pedestrian steps into the road);
- **d.** the hazard was caused by other road users (rather than by the behaviour of the film car driver); and
- **e.** at the point of hazard onset, other hazard precursors had to be present to provide plausible distracter options for the hazard prediction version of the test.

Figure 2.1: Graphic representation of the location of cameras attached to the film car (the blue dot represents the positioning of the 360-degree camera and the red dots represent the cameras used to capture the mirror information)



Source: Megan Choud, The Noun Project, modified by the authors

Table 2.2: Description of the 12 hazards that comprise the two tests (with correct answers underlined) created for Study 1

Clip number	Description	Multiple-choice options (with correct answers underlined)	Full clip length (sec)	Hazard onset (sec)	Hazard offset (sec)	Occlusion point (sec)
1	While you are 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.	1. An oncoming car encroaches on your lane. 2. A car pulls out in front of you from the side road on the left. 3. A pedestrian steps out into the road from the left. 4. A pedestrian runs out into the road from the bus stop on the right.	51	41.3	45.1	42.1
2	While you are 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.	1. A car reverses out of the side road on the right. 2. The car ahead brakes suddenly. 3. A parked car on the left pulls off in front of you. 4. The driver's door of the parked car ahead opens.	54	35.5	40.8	39.3
3	You are driving along a two-lane road, when a car stopped in the left lane ahead causes you to change lanes to avoid a collision.	1. An oncoming car turns across your path into the side road on the left. 2. A cyclist emerges from the side road on the left. 3. A parked car blocks your lane ahead. 4. A car overtakes you from the right.	29	24.3	27.9	26.0
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.	1. A pedestrian steps into the road from the left. 2. A van pulls out in front of you from the side road on the right. 3. A parked car on the left pulls off in front of you. 4. The car ahead brakes suddenly for the pedestrian crossing ahead.	37	23.3	27.4	23.8
5	You are travelling down a busy town centre high street when a car pulls off on the left, blocking your path. As you slow down, two pedestrians use this to cross the road in front of you.	1. Pedestrians step out into the road from the right. 2. A parked car on the left pulls off in front of you. 3. The driver's door of the car ahead suddenly opens. 4. Pedestrians step out into the road from the left.	43	25.2	33.3	25.5

Clip number	Description	Multiple-choice options (with correct answers underlined)	Full clip length (sec)	Hazard onset (sec)	Hazard offset (sec)	Occlusion point (sec)
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.	1. A car emerges from the side road on the left. 2. A pedestrian steps from the right into the road at the crossing. 3. An oncoming car turns across your path into a side road on the left. 4. A car undertakes you at speed.	57	45.2	47.9	46.1
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.	1. The van in the adjacent lane to the right pulls in front of you. 2. A pedestrian crosses the road from the right. 3. A pedestrian steps out into the road from the left. 4. A car pulls out from the left, blocking your path.	37	22.8	27.9	23.5
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.	1. A parked van pulls into the road from the right blocking your path. 2. A road worker steps into the road from the right. 3. An oncoming van approaches in your lane. 4. A fallen road sign blocks your lane.	38	22.1	27.1	23.8
9	You are queuing 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.	1. The car in the right lane indicates and pulls into your lane. 2. A car pulls out from the side road on the right. 3. A pedestrian gets out of the car ahead. 4. A motorcyclist overtakes you from the right.	25	16.4	25.0	16.8
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.	1. A pedestrian steps into the road from the left. 2. A parked lorry blocks your lane. 3. A white van overtakes you from the right. 4. A car pulls out in front of you from the side road on the left.	28	15.9	21.9	20.1

Clip number	Description	Multiple-choice options (with correct answers underlined)	Full clip length (sec)	Hazard onset (sec)	Hazard offset (sec)	Occlusion point (sec)
11	You are driving along a heavily congested road. A delivery van ahead 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.	1. A pedestrian steps out into the road from the left. 2. A pedestrian with a pushchair appears in the road from behind the turning van. 3. The van ahead brakes suddenly to avoid pedestrians stepping out into the side road. 4. The parked pulls into the road from the left blocking your path.	27	15.2	18.1	15.9
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.	1. A pedestrian steps out from behind the parked car on the left. 2. The car ahead performs a U-turn in the road. 3. The white car parked on the left pulls out in front of you. 4. The car ahead reverses towards you.	39	18.9	27.0	25.1

Source: Authors' own (Study 1)

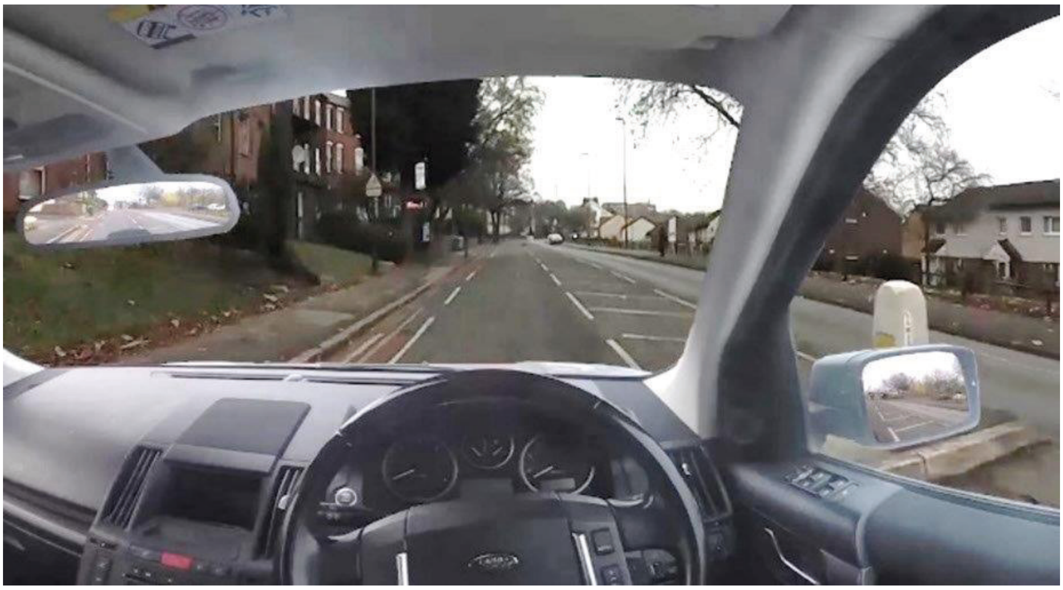
Following selection, the multiple video feeds were synchronised. The 360-degree view was wrapped around a 360-degree photograph of the interior of a Land Rover Freelander (chosen to fit with the viewing position afforded by the 360-degree footage as recorded from the roof of the Vauxhall Corsa). Video from the rear-facing GoPro cameras was edited into the mirror placeholders of the Land Rover. The result was an immersive video viewed as if from the driver's perspective (Figure 2.2).

The 12 hazard clips were edited to create both the hazard perception and hazard prediction formats. For the perception format, the team of traffic psychologists agreed on temporal scoring windows for each clip, starting at the onset of the hazard, and terminating at its offset (the time at which the hazard ends). The scoring window was then divided into five segments to create five scoring zones, for the awarding of points from five to one for any response made within the window (Figure 2.3).

For the hazard prediction test, an occlusion point was chosen for each of the 12 clips. This was the point at which the clip would suddenly end and cut to a black screen containing the question "What happens next?" The occlusion point is typically at or just after the point of hazard onset: if the viewer has correctly predicted that a pedestrian may step out from behind a parked car, it is highly likely that they will be looking at the parked car just as the pedestrian begins to emerge.

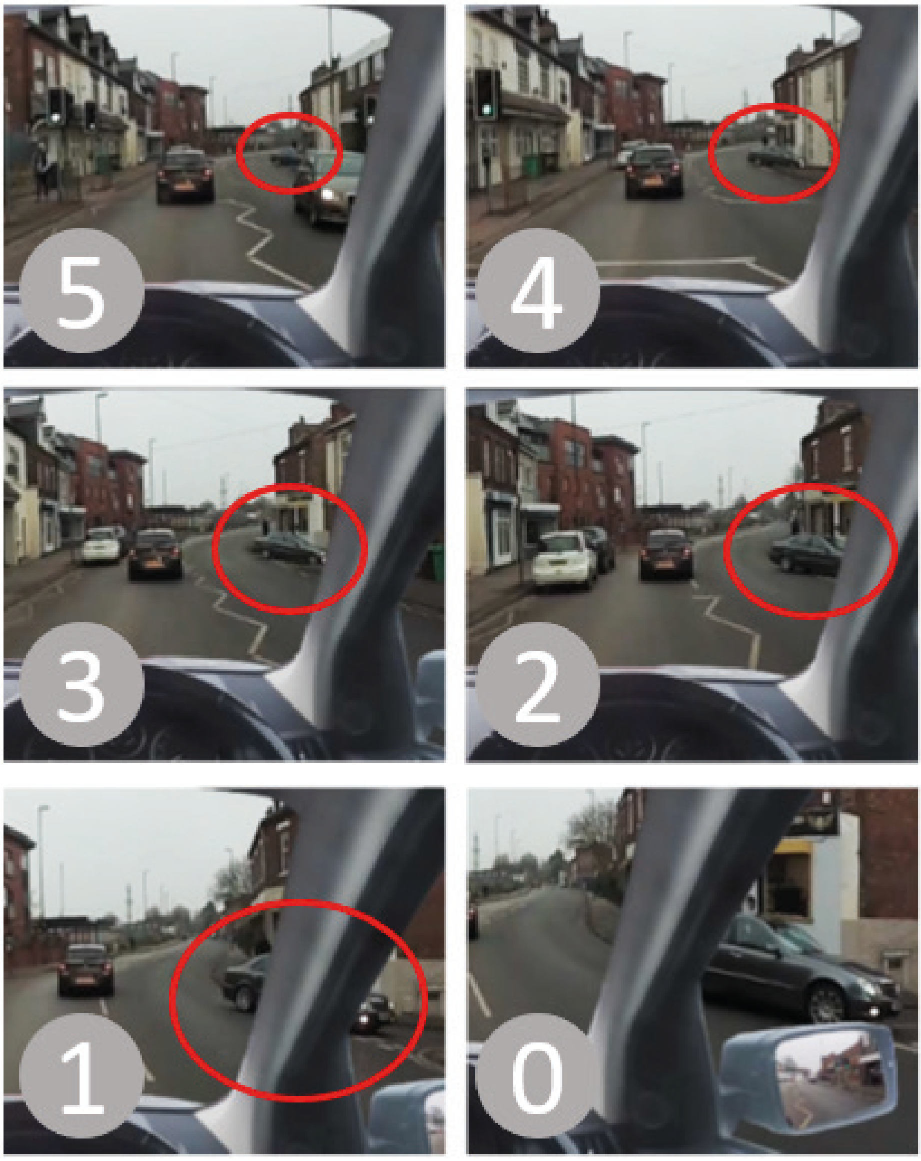
A glimpse of the actual hazard that is shown between onset and occlusion should be enough to confirm that their prediction was correct. Crucially, the occlusion point should not be so late that the hazard fully materialises and can attract the attention of all viewers regardless of whether they predicted it or not. Following selection of the occlusion points, the research team created four options for each clip: one correct answer and three plausible, but incorrect, distracters. The team debated the plausibility of each option, and the precise wording. The questions and options were edited on to the end of each clip, with the target appearing in a randomly determined location within the list of four options (Figure 2.4).

Figure 2.2: Screenshot taken from the 360-degree immersive video footage created for the Study 1 tests



Source: Authors' own (Study 1)

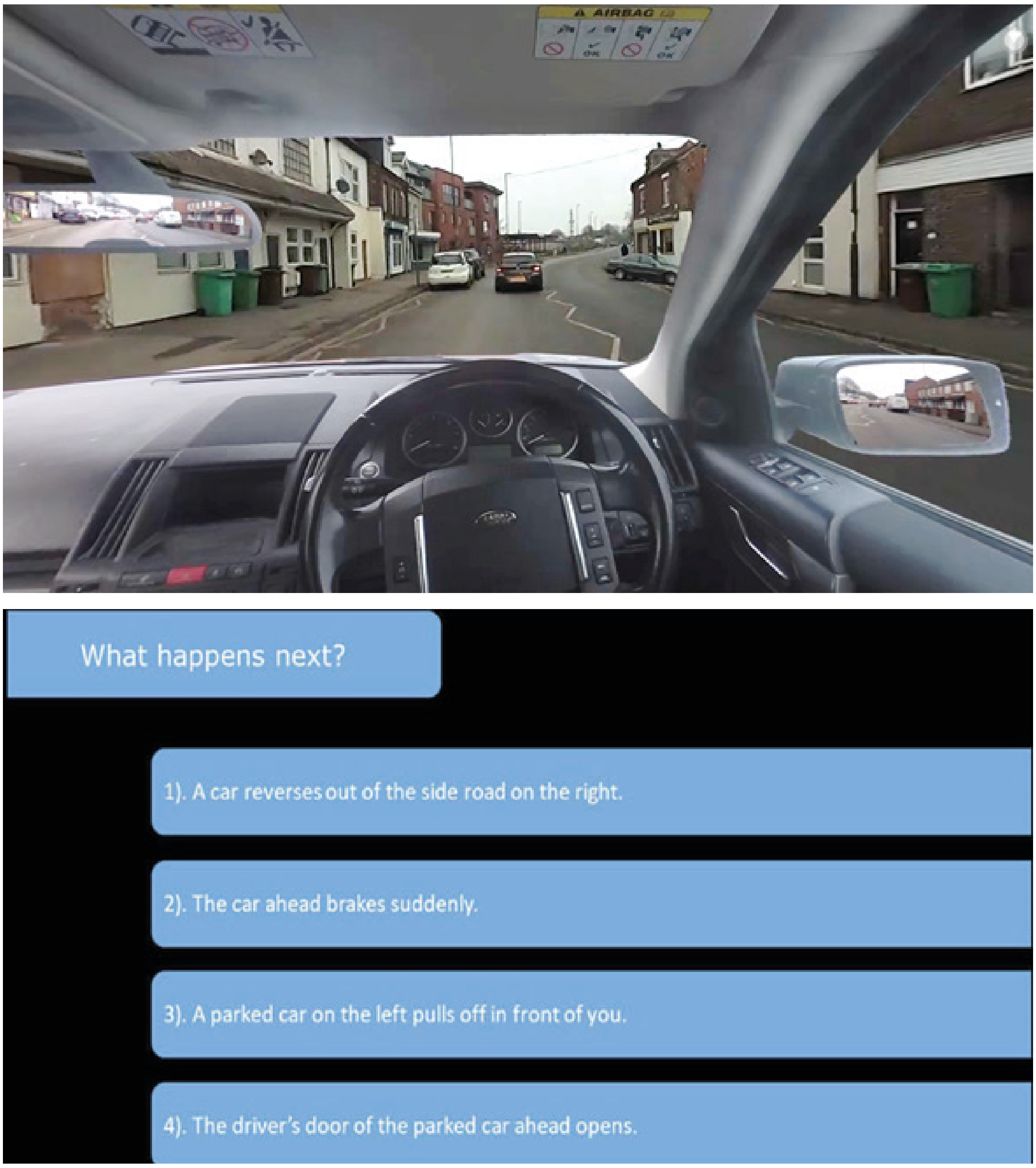
Figure 2.3: Images depicting the start of the scoring windows for Hazard 2 (Table 2.2)



Source: Authors' own (Study 1)

Note: The red circles are not present in the clips when viewed by participants and are merely used here to denote the location of the hazard; the numbers represent the points associated with the relevant portion of the scoring window

Figure 2.4: A hazard prediction clip typically plays up to the point where the hazard has begun to materialise (top panel), after which the screen is occluded and participants are asked "What happens next?" (bottom panel) – they respond by choosing from four options available on the screen



What happens next?

1). A car reverses out of the side road on the right.

2). The car ahead brakes suddenly.

3). A parked car on the left pulls off in front of you.

4). The driver's door of the parked car ahead opens.

Source: Authors' own (Study 1)

The average clip length of the hazard perception clips was 38.7 seconds (range: 24 s to 57 s) and the mean hazard onset was 25.6 seconds (range: 15.2 s to 45.2 s, Table 2.2). Hazard prediction clips were shorter, typically occluding just after the point of hazard onset.

2.2.3.2 The practice clip

Before viewing either of the tests, participants were acclimatised to the virtual environment through exposure to a practice clip. This clip was adapted from a 360-degree road safety video produced by Road Safety Scotland.

Two versions were created from the Road Safety Scotland footage: a short, initial practice clip (40 seconds long) and a longer, more immersive clip that was 2 minutes and 13 seconds in length.[7] This clip showed the driver's perspective of a test drive, which culminated in a single unexpected hazard (a pedestrian crossing the road). Participants were not required to make any response to this clip.

2.2.3.3 The questionnaires

All participants who completed the study were asked to complete the following questionnaires:

Demographics questionnaire: in addition to age and gender, this questionnaire included questions to assess annual mileage, hours of driving per week, years of experience, collisions (number, severity, blame), and violation points.

The simulator sickness questionnaire: 16 factors (symptoms) assess participants' sickness levels (e.g. sweating, nausea, fatigue etc.). This assessment was made at five time points throughout the testing procedure (see subsection 2.2.5). Participants rated each symptom on a four-point scale (from none to severe). Testing was aborted if:

- any single factor was reported as 'severe' (even at time point 1);
- any single factor increased by two stages between time point 1 and time point N – e.g. 'sweating' at time point 1 is recorded as 'none', but at time point 2 is recorded as 'moderate'; or
- three or more factors increased by one stage from time point 1 to time point N.

CRIE questionnaire: participants were asked to rate each test on four Likert scales:

- *Comfort* – How comfortable did you feel during the experiment? Responses were on a ten-point scale from 1 (extremely uncomfortable) to 10 (extremely comfortable).
- *Realism* – How realistic was the task? E.g. How close to real life was it? The rating scale ranged from 1 (extremely unrealistic) to 10 (extremely realistic).
- *Immersive* – How immersive was the task? e.g. did it feel like you were there? The rating scale ranged from 1 (extremely un-immersive) to 10 (extremely immersive).
- *Engagement* – How engaged did you feel with the task? (e.g. looking where a driver would look). The rating scale ranged from 1 (extremely unengaging) to 10 (extremely engaging).

2.2.4 Apparatus

An HTC Vive headset with a Tobii eye tracker (2,160 × 1,080 resolution, with a sample rate of 120 Hz) was used to display both tests. A Republic of Gamers ASUS ROG Strix Hero III Gaming Laptop was used to administer the tests using Tobii Pro Lab software to design and present the experiments. A keyboard was given to participants to press when they saw hazards in the hazard perception test.

[7] www.youtube.com/watch?v=hnWgEGVjlak

2.2.5 Procedure

Participants were invited to the laboratory at Nottingham Trent University. Upon arrival, participants were given instructions and asked to sign a consent form followed by a demographics questionnaire to collect age, gender, and driving history data. The experimental protocol was as follows:

- *Simulator sickness checklist* (time point 1) to gauge their baseline sickness symptoms
- A short practice clip (40 seconds of the Road Safety Scotland video)
- *Simulator sickness checklist* (time point 2)
- A long practice (2 minutes 13 seconds of the Road Safety Scotland video)
- *Simulator sickness checklist* (time point 3)
- Hazard test 1 (either perception or prediction test, counterbalanced across participants)
- *Simulator sickness checklist* (time point 4) and CRIE questions
- Hazard test 2 (either perception or prediction test, counterbalanced across participants)
- *Simulator sickness checklist* (time point 5) and CRIE questions

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 presented in 360 degrees, taken from the perspective of a driver.

For the hazard perception test, participants were given a keyboard to rest on their laps. They were instructed to press a key on the 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. For the hazard prediction test, drivers were instructed to watch the clips and search for potential hazards. They were told that the clip would stop suddenly, and the image would be occluded just as a hazard begins to unfold. Following this, they were presented with four possible options (numbered 1–4) regarding what might happen next, from which they had to select the correct answer. They verbally reported their option to the experimenter, who entered their response via the controller keyboard. It was done this was because participants would not be able to view any keyboard given to them to select their answer without taking off the VR headset, which would disrupt the calibration and immersion. The testing session lasted around 30–45 minutes, for which participants received a £10 voucher.

2.3 Results

Four participants (5.3%; one in the 17–25 age group, two participants in age group 46–55 and one in the 56+ group) were removed because their sickness ratings rose above threshold (and a further participant was removed as a result of data loss). The demographics of the four drivers removed on the grounds of cybersickness symptoms are shown in Table 2.3.

Table 2.3: Demographics of the four participants removed from Study 1 owing to their simulator sickness questionnaire scores

Group	N	Gender	Mean Age (years)	Mean Driving experience (years since passing test)
17–25	1	1 female	21.4	4.4
46–55	2	2 females	52.6 & 51.2	25 & 30
56+	1	1 female	58.8	30.8

Source: Authors own (Study 1)

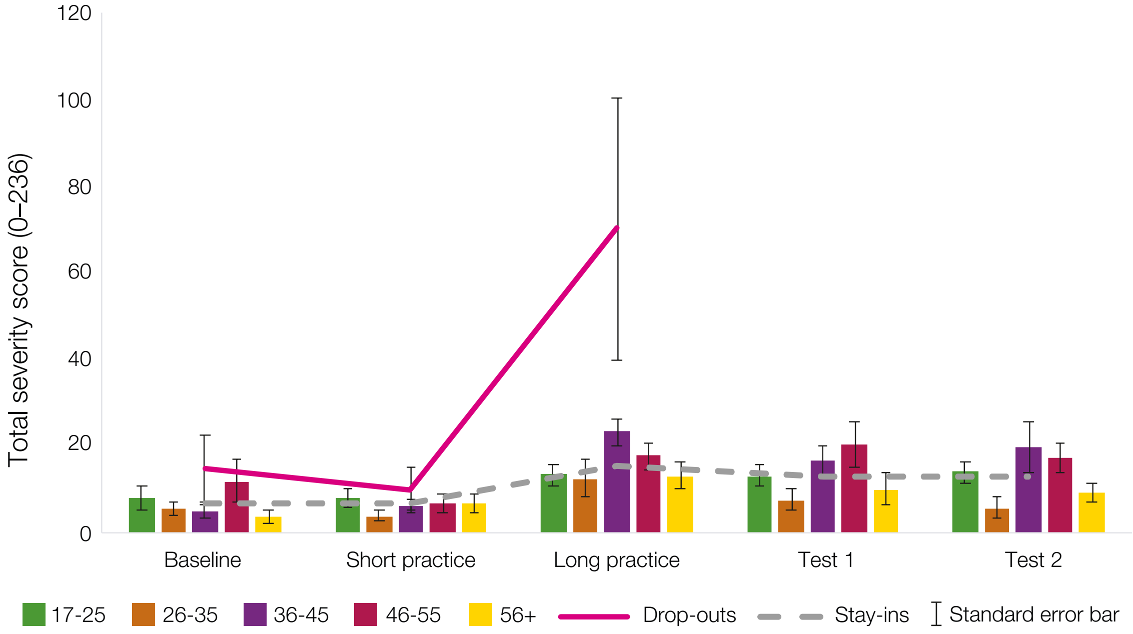
2.3.1 Sickness ratings over time

Participants' reported sickness severity was calculated for the five points at which the SSQ was administered by using the method explained in Kennedy et al. (1993). This requires individual item scores (0, 1, 2, or 3) to be summed within three subscales for each participant (nausea, ocular discomfort and disorientation; with some items loading on two subscales). 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 in the range 10–15 reflect significant sickness symptoms, scores in the range 15–20 indicate more serious issues, and scores over 20 suggest that there is a problem that, until rectified, will probably prevent the simulator from being used.

As can be seen from Figure 2.5, the four participants who were withdrawn because of sickness were identified at time point 3, following presentation of the full practice clip (2 minutes, 13 seconds). All participants who did not exhibit above-threshold sickness levels after the practice clip went on to complete the rest of the experiment. This suggests that exposure to 360-degree driving footage lasting just over two minutes may be sufficient to identify those participants who should probably not continue.

These scores were subjected to a 5 × 5 mixed analysis of variance (ANOVA) (time of questionnaire × age group, not including the four 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 revealed that the only difference between the times was a significant increase in sickness ratings following the long practice on the ratings given after the short practice (even after Greenhouse-Geisser and Bonferroni corrections: $F(1, 67) = 40.7$, $MSE = 150.0$, $p < .001$, $\eta_p^2 = .38$). Neither the main effect of age, nor the interaction between age and the time when the SSQ was administered, were significant.

Figure 2.5: Participants' sickness ratings in Study 1 at each time point for each driver age group)



Source: Authors' own (Study 1)
Note: bars represent the group means for different ages; dashed line represents the group mean across all ages; red line represents the mean of the four participants who were removed

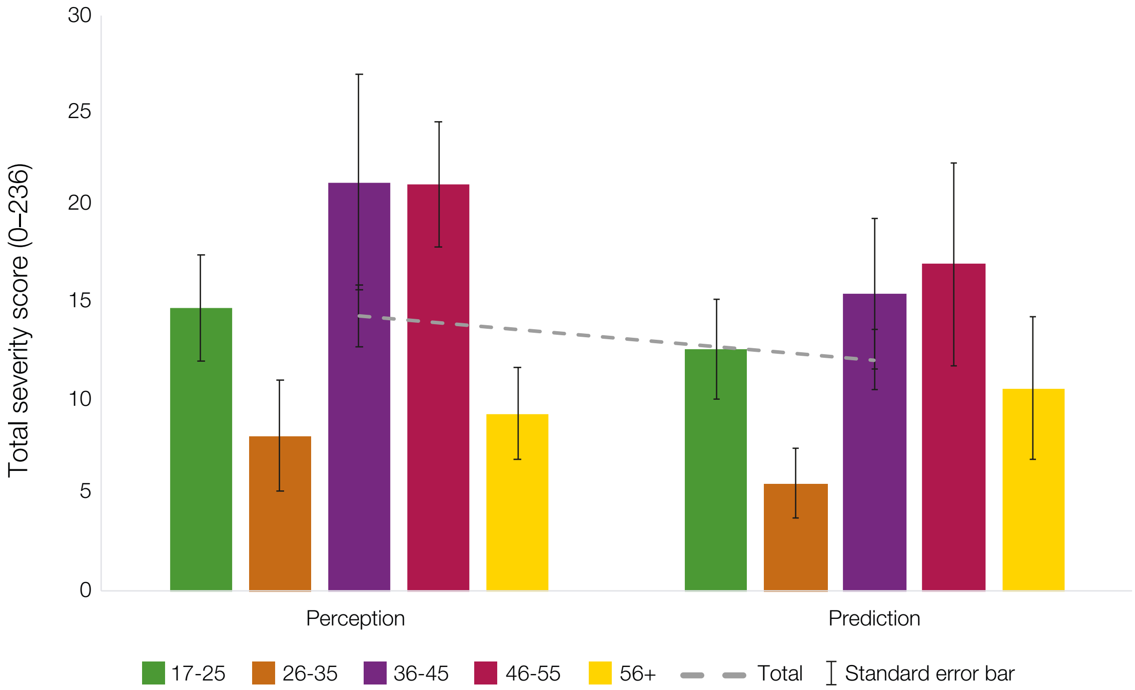
2.3.2 Sickness ratings for hazard prediction and hazard perception

The data in Figure 2.5 shows the sickness scores for the two tests as they were presented in order. Test 1 was the perception test for half of the participants, and the prediction test for the other half of the participants (and vice versa for Test 2). For the next analysis the data were recategorised according to the format of the test (prediction or perception), and then compared. The data were compared by means of a mixed 2 × 5 ANOVA (test type × age group). This analysis revealed a main effect of test type ($F(1, 67) = 5.2$, $MSE = 44.7$, $p = .03$, $\eta_p^2 = .08$), revealing that participants in the hazard perception test had significantly higher mean sickness levels than those in the hazard prediction test (14.3 vs 12.1). The interaction approached significance, but did not cross the threshold, $F(4, 67) = 2.3$, $MSE = 44.7$, $p = .06$ As can be seen in Figure 2.6, following the hazard perception test, two of the age groups (36–45 and 46–55) produced SSQ ratings that crossed Kennedy et al.'s (2003) suggested boundary value of 20 for identifying a serious cybersickness problem with the test. SSQ scores for these age groups following the hazard prediction test were, however, below this threshold, but nevertheless hovered around the point at which symptoms might be considered a cause for concern.

2.3.3 Comfort, realism, immersion and engagement questions

Participants gave ratings for each of the questions regarding CRIE on a ten-point scale for both the hazard perception and the hazard prediction tests. Each rating was entered into a mixed 2 × 5 ANOVA (test type × age group).

Figure 2.6: The total severity score of the sickness ratings for the hazard perception and prediction tests of Study 1



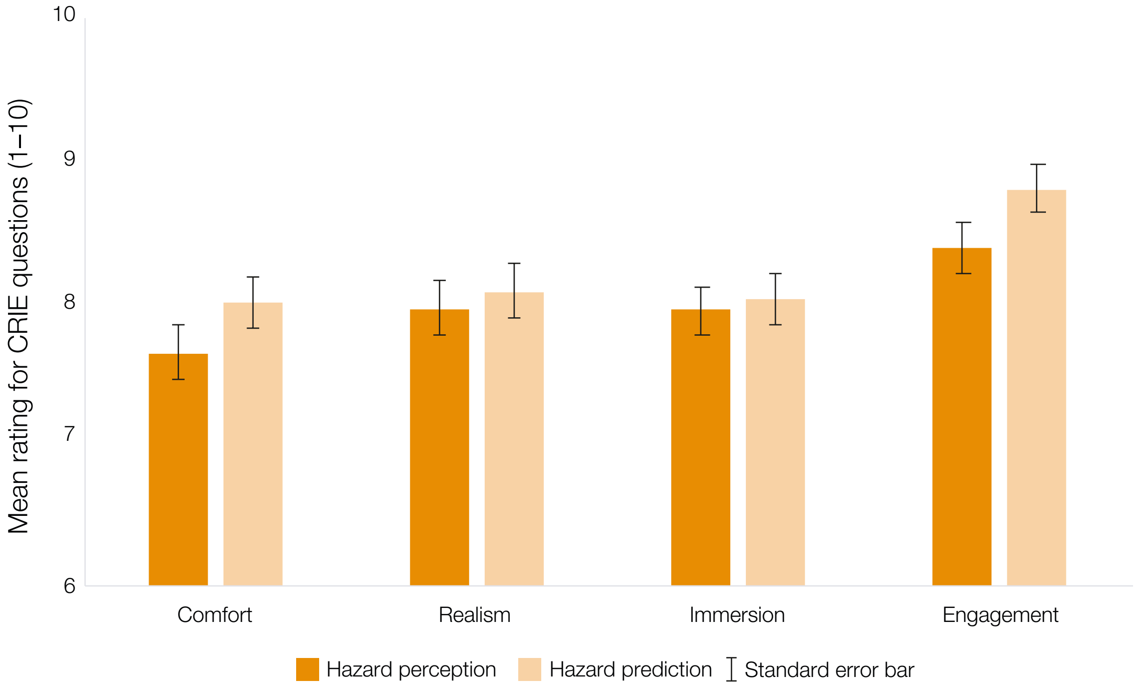
Source: Authors' own (Study 1)

Main effects of test type were found for comfort and engagement ratings, with participants rating the hazard prediction test as more comfortable than the perception test (with mean ratings of 8.0 and 7.6; (F(1, 67) = 6.4, MSE = 0.7, p = .01, η_p^2 = .08)), and also more engaging (8.8 vs 8.3; F(1, 67) = 10.7, MSE = 0.8, p = .002, η_p^2 = .14). The mean CRIE ratings are displayed in Figure 2.7.

Comfort ratings were also affected by age (F(4, 67) = 2.8, MSE = 3.9, p = .03, η_p^2 = .14), suggesting that participants aged 46–55 (M = 7.1) and 55+ (M =7.4) found the tests to be less comfortable than those participants in the 26–35 (M = 8.9) age group (Figure 2.8). Evidence for an interaction between test type and age group for comfort ratings approached the threshold for significance, F(1, 67) = 2.4, MSE = 1.8, p = .059, η_p^2 = .13. Figure 2.8 suggests that it was the older participants (36+) who reported the prediction test to be more comfortable. There were no other significant main effects or interactions for any of the CRIE questions (all values of p > .05).

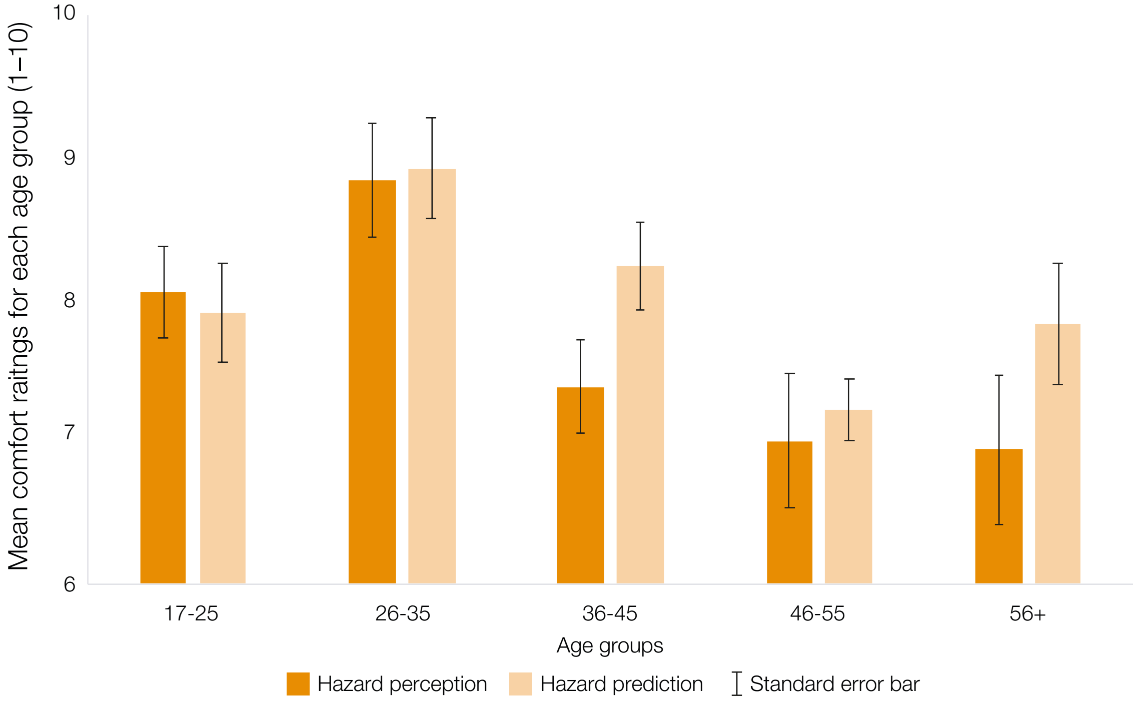
Hierarchical regressions were undertaken to assess the impact of age and CRIE ratings on sickness scores for the two tests. Participant age was entered at stage 1 of each regression, followed by CRIE ratings at stage 2 (producing two models, see Table 2.4). The regression model for sickness scores on the hazard *perception* test revealed that at stage 1, participant age did not contribute significantly to the regression model, *F*(1, 70) = 0.2, *p* = .64, and accounted for only 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* = .11.

Figure 2.7: Average ratings given for each of the four items on the Study 1 CRIE questionnaire for all age groups



Source: Authors' own (Study 1)
Note: *: p = .01; **: p = .002

Figure 2.8: Average comfort ratings given for each age group, Study 1



Source: Authors' own (Study 1)

When this analysis was repeated for the *hazard prediction* sickness scores, age once again did not predict sickness, *F*(1, 70) = 0.1, *p* = .77, and accounted for only 0.1% 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, and this change in R^2 provided a marginally significant result (*F*(5, 70) = 2.4, *p* = .05). However, of the four CRIE questions, the only items that influenced hazard prediction sickness were comfort (β = –0.43, *t*(65) = –3.2, *p* = .002) and engagement (β = 0.37, *t*(65) = 2.1, *p* = .04). While increased comfort results in a decrease in sickness symptoms, the relationship between engagement and sickness is surprising in that it positively relates to an increase in sickness symptoms.

Table 2.4: Factors influencing participant sickness scores for the hazard perception test and the hazard prediction test in Study 1

	Dependent variable	Independent variable	*Unstandardised coefficients* (B, standard error)		Standardised coefficients (β)	t-value		R² (adjusted R²)	F-value	P value	
Hazard perception sickness	**Model 1**	Age	–0.41	0.09	–0.06	–0.47		0.003 (–0.01)	0.23	.64	
	Model 2	Age	–0.12	0.09	–0.16	–1.32		0.13 (0.06)	1.87	.11	
		Comfort	–3.01	1.11	–0.37	–2.71					
		Realism	–1.81	1.43	–0.22	–1.27					
		Immersion	0.97	1.79	0.10	0.54					
		Engagement	0.95	1.10	0.11	0.87					
Hazard perception sickness	**Model 1**	Age	0.02	0.08	0.04	0.29		0.001 (–0.01)	0.09	.77	
	Model 2	Age	0.03	0.08	0.05	0.39		0.15 (.09)	2.35	.05	*
		Comfort	–3.98	1.25	–0.432	–3.18	**				
		Realism	-0.33	1.48	–0.04	–0.22					
		Immersion	0.20	1.62	0.20	0.12					
		Engagement	3.47	1.65	0.37	2.10	***				

Source: Authors' own (Study 1)

Note: One participant (age group 55+) was excluded from this analysis for failing to provide an exact date of birth resulting in N = 71 for both regressions. 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, $\Delta R^2 = 0.152$; *: $p = .05$; **: $p = .002$; ***: $p = .04$.

2.3.4 Hazard perception and hazard prediction performance

Hazard perception or prediction performance on the two tests was not a primary consideration in the current study as we had relatively few clips per test (N = 6), and the study was not designed to compare high-risk and low-risk (or experienced and inexperienced) drivers. However, for completeness, the performance of the sample is given in Table 2.5. For this analysis, one participant (age group 26–35) was removed owing to equipment failure resulting in a loss of their hazard perception data. Mixed ANOVAs comparing performance on the two tests across driver age group were conducted on the data; however, these did not reveal any significant main effects or interactions (all values of $p > .05$).

Table 2.5: Participants' performance on both the hazard tests of Study 1 across the age groups

	Hazard perception			Hazard prediction	
Group	**N**	**Score (0–30)**	**Number correctly identified (0–6)**	**N**	**Number correctly predicted (0–6)**
17–25	21	14.4	4.0	21	4.6
Comprising young, moderately experienced drivers	*(19)*	*(14.9)*	*(4.1)*	*(19)*	*(4.7)*
plus one learner	*(1)*	*(18)*	*(6)*	*(1)*	*(4)*
plus one novice	*(1)*	*(1)*	*(1)*	*(1)*	*(4)*
26–35	11*	14.8	4.1	12	5.0
36–45	13	17.1	4.7	13	4.5
46–55	9	15.8	4.6	9	5.0
56+	17	16.4	4.5	17	4.6
All participants	**71**	**15.6**	**4.3**	**72**	**4.7**

Source: Authors' own (Study 1)

Notes: (a) The youngest age group is broken down to reveal the performance of our learner and novice drivers. The four participants who were removed with cybersickness did not undertake either test. (b) *Equipment failure led to the loss of hazard perception behavioural data for one participant in this condition.

2.3.5 Analyses of eye movements

As noted above with the behavioural data, eye-movement measures were not a primary concern for this study, though they are included for completeness. A series of 2 × 5 mixed ANOVAs compared a selection of participants' eye-movement measures across age groups and test variants. These measures included the number of hazardous precursors that participants looked at, how quickly they looked at the precursor once it was visible ('time to first fixate'), and dwell time on the hazardous precursors prior to occlusion (as a percentage of the time each precursor was visible). For these analyses, data from three participants from the hazard perception test (age groups 26–35, 46–55, 56+), one participant from the hazard prediction test (age group 36–45), and one participant from both tests (age group 17–25) was removed owing to poor calibration. None of these measures showed any significant main effects or interactions (all values of $p > 0.05$). The means for these measures can be found in Table 2.6.

Table 2.6: Participants' eye-movement measures across the two tests and five age groups in Study 1 (excluding four participants removed with cybersickness who did not undertake either test)

	Hazard perception				Hazard prediction			
Group	**N**	**Did they look? (0–6)**	**Time to first fixate (0–5)***	**Dwell time (%)**	**N**	**Did they look? (0–6)**	**Time to first fixate (0–5)**	**Dwell time (%)**
17–25	20	5.6	3.7	50.8	20	4.4	2.7	35.7
26–35	11	5.3	3.7	48.7	12	5.1	3.2	44.1
36–45	13	5.3	3.5	40.9	12	4.8	2.9	41.7
46–55	8	5.9	4.1	56.9	9	4.7	2.8	41.3
56+	16	5.8	3.8	49.8	17	4.8	2.6	36.5
All participants	**68**	**5.6**	3.8	49.3	70	4.7	2.8	39.1

Source: Authors' own (Study 1)

Note: * Time to first fixate is calculated in the same manner as the DVSA scoring method for hazard perception, with points awarded for how quickly one first fixates in the hazard precursor window (see subsection 3.3.3.2).

2.4 Discussion

This study aimed to assess the levels of cybersickness induced by 360-degree hazard tests when presented in a VR headset. A specific aim was to compare a traditional speeded-response hazard perception test with an occlusion-based hazard prediction test to establish the best test format to use for the remaining studies in this project. The results demonstrated that, contrary to our concerns, the hazard prediction test evoked significantly less-severe sickness symptoms and was rated as more comfortable and engaging than the hazard perception test. Behavioural measures and eye movements revealed no differences between the tests or age groups, but given the small number of intended clips that were designed for this first study, we did not anticipate any findings of interest (nor did we seek to recruit groups of drivers that were likely to differ on these measures).

Overall, the results showed a relatively low number of people suffered from excessive symptoms of sickness. A total of four participants were removed from the study following the long practice clip, with an average score of 70 on the SSQ, which is extreme compared to those of the other participants. Two of these four participants explicitly dismissed the idea of becoming nauseous prior to the study and were surprised with the onset of significant symptoms. While these participants expressed regret at being unable to continue, all four were happy to be withdrawn from the study owing to the severity of cybersickness.

During debriefing, the four participants reported that their symptoms increased throughout the long practice clip, though none of them asked to withdraw. This raises the importance of repeated monitoring of sickness symptoms: despite awareness of their right to withdraw at any point without explanation, it is likely that they would have tried to continue with the study if the experimenter had not enquired after their wellness.

Even the participants who were not removed from the study reported a significant increase in sickness symptoms following the long practice. Fortunately, their rise in symptom severity was not sufficient for the experimenter to withdraw them, and these symptoms then plateaued throughout the hazard perception and hazard prediction tests. None of these participants reported the highest level on any of the symptom scales, and there were only 42 instances where a participant chose the third position on the 1–4 scale (0.7% of all responses). Furthermore, 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 distributed across the age groups. There was, however, 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 inexperience with VR headsets or with hazard tests.

2.4.1 How nauseating are the current tests compared to the literature?

Before discussing the comparison of the sickness scores evoked by the two tests, it is worthwhile comparing the overall levels of sickness found in this study with levels reported in the general literature. Kennedy et al. (2003) suggested that mean SSQ ratings between 10 and 15 should be considered significant symptoms. Scores between 15 and 20 indicate more serious issues, while scores above 20 suggest fundamental problems with the system (see subsection 2.3.1). As the current participants reported a mean of 6.6 on this scale at baseline, it was always likely that symptoms would rise above Kennedy et al.'s threshold for problematic simulator sickness. For those participants who remained in the study following the long practice, their sickness scores averaged 15.3, which then dipped slightly to approximately 13 across the hazard tests. According to Kennedy et al.'s guidelines, our tests cause significant levels of sickness in absolute terms.

The situation is less bleak, however, if we compare our sickness ratings with more recent studies. For instance, Saredakis et al.'s (2020) meta-analysis of 55 VR studies found an average sickness score of 28.0, ranging from 14.3 to 35.2. Saredakis et al. suggested that the observed differences between their review and the scores provided by Kennedy et al. could be due to Kennedy's use of military pilots. These pilots may have been less affected by sickness owing to flight experience and training (or perhaps they were simply more motivated not to report it). Furthermore, the level of visual detail involved in driving around a bend in a car is of a magnitude greater than a similar turn in a plane. Arguably, it is unfair to compare sickness symptoms in a driving simulator to those of a flight simulator.

The average sickness ratings for the current tests fall at the lower end of the range of scores reported in the 55 studies reviewed by Saredakis et al. Even those age groups who reported the highest level of cybersickness symptoms (36–45, 46–55) still gave ratings considerably below the mean sickness ratings found in the meta-analysis (<28). If we added the four participants were removed because of sickness (assuming their scores of 70 plateaued across the tests) this would still leave the current mean sickness scores in the lowest quintile of the range provided by Saredakis et al. Thus, compared to a wider and more relevant evidence base, the current tests fare very well in regard to the levels of induced cybersickness.

Our relatively low sickness ratings could be explained by several factors, including the length of the tests. Each test lasted no longer than five minutes. Including the practice clip, participants were immersed for approximately 15 minutes in total, and took several breaks from the headset between the practices and the tests. This is a much shorter time in a virtual environment than other studies with higher dropout rates (Saredakis et al., 2020). Another potential reason for our lower sickness levels 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, 2004b). One theory of simulator sickness suggests that symptoms are not necessarily evoked by mismatched visual and vestibular motion cues, but are instead caused by conflict with the so-called 'rest frames', which are parts of the virtual environment that are consistent with the real world (Prothero, 1998; Prothero & Parker, 2003). In this instance, the graphic overlay of the car interior may have provided an independent background with which the participants could orient themselves, thus reducing the overall mismatch between the virtual and physical environment.

2.4.2 Age and cybersickness

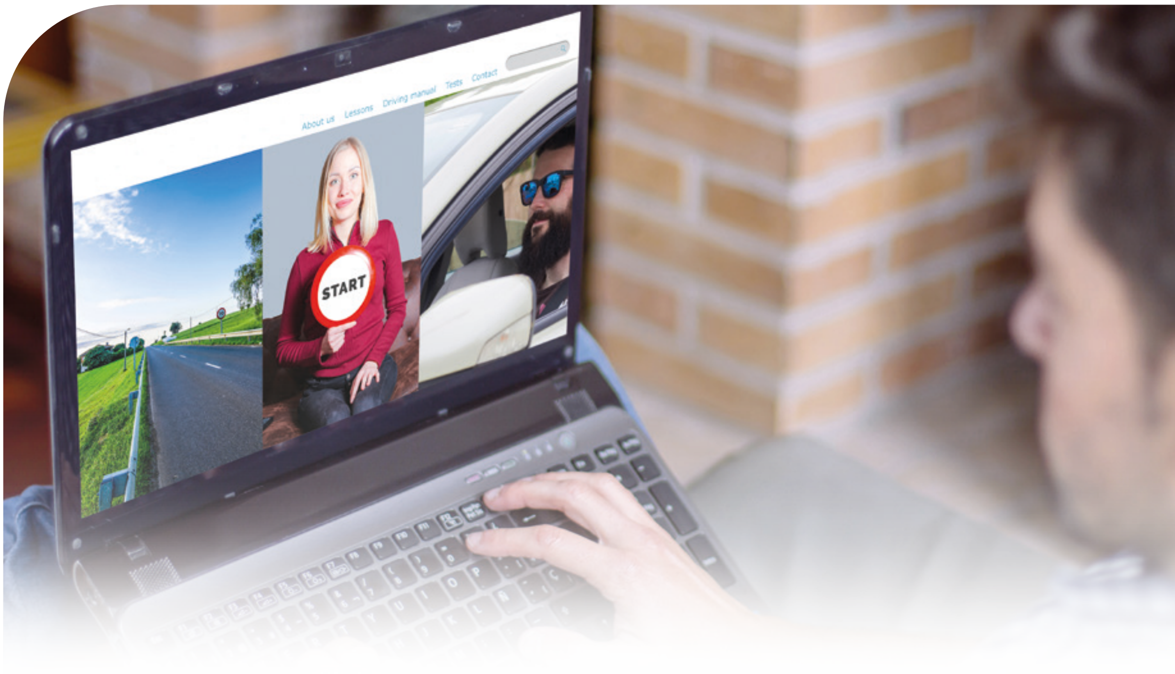
As noted in the introduction, the literature paints an unclear relationship between cybersickness and age (Kennedy et al., 2010; Benoit et al., 2015). While some argue that age is positively related to sickness symptoms (Cassavaugh et al.,2011; Classen et al., 2011; Golding 2006; Matas et al., 2015; Trick & Caird, 2011), Saredakis et al. (2020) suggested that their meta-analysis indicates that such a relationship may have been overstated (though they recognise that this conclusion is based on a relatively small number of studies with older participants). Unfortunately, the current study does not provide a clear steer in this debate. Though symptom severity did appear to vary across ages groups, there was no consistent or significant effect. There was, however, a significant difference in participants' enjoyment of the tests, with older participants reporting lower levels of comfort than younger drivers. While this finding might not manifest as a need to withdraw older participants from VR studies, it suggests that older drivers might be less willing to accept VR as a training or assessment method outside of the study environment.

2.4.3 The comparison of hazard perception and prediction tests

Despite concerns that the occlusion technique of the hazard prediction test would induce more severe sickness symptoms, the opposite was found to be true. Why might this be the case? One possible explanation could relate to the slightly shorter duration of the prediction clips (over ten seconds shorter on average, owing to the occlusion occurring at hazard onset). This need not be considered a confound in the comparison of perception and prediction clips, but a genuine advantage of the prediction test. Alternatively, the opportunity to read the multiple-choice options in the prediction test may have provided a sufficient break for symptoms to abate slightly. This could explain why increased comfort ratings were associated with reduced sickness in the regression. It was, however, surprising to find that increased engagement led to greater sickness. While this detrimental relationship only just passed the threshold of significance, it is worth following up in future studies.

In conclusion, both tests fare relatively well in terms of overall sickness rates. While we feared that the prediction test might evoke higher levels of sickness as a result of the sudden occlusions, in fact, the opposite was the case. The hazard prediction test was preferred by participants, being rated as more comfortable and engaging. Given the previously noted benefits of the hazard prediction test over the perception test, such as a fairer and more transparent scoring system, the current findings support the use of the hazard prediction test format for the remaining studies in this project.

3. Study 2: A Comparison of a Video-Based 360-Degree Hazard Test and a Single-Screen Hazard Test



3.1 Introduction

The second study was designed to address the primary concern of this project: is it preferable to measure hazard awareness by means of 360-degree clips presented in VR headsets, or by using a more traditional presentation method via a single computer monitor? The answer to this question is, however, dependent on how the success of a VR-based hazard test is defined.

First, one might consider the preferences and experiences of drivers who are likely to engage in such tests. An obvious cohort to target is learner drivers who must pass the UK hazard perception test before they are allowed to take their on-road test.

The ‘threat’ of the hazard perception test encourages many young drivers to seek out hazard perception training materials from commercial providers. These drivers are likely to be a primary market for VR hazard training. Increasingly, however, professional drivers are also being exposed to VR-based driver training materials as part of government-mandated training hours (35 hours training per five years). One recent example is a Transport for London initiative which required VR training to be designed for roll-out to 25,000 London bus drivers (Destination Zero, 2019–2021[8]). The experiences of these driver groups after engaging in VR-based assessment and training is an important dependent variable that must be considered. If drivers do not feel comfortable throughout the experience, this will probably affect their future preferences. Equally, if the experience is jarring or distracting, or simply does not feel realistic, then this may also have a negative impact on future acceptance of such technology. The CRIE questions demonstrated promise as indicators of drivers’ personal experience in Study 1. For this reason, they were employed in Study 2 also. CRIE ratings following a VR hazard test are likely to differ to those recorded after a single-screen test. It is unlikely that engaging with a VR test will be considered more ‘comfortable’ than undertaking a single-screen test, though we hoped that positive benefits in terms of realism, immersion and engagement might offset any negative comparisons.

In addition to participants’ experiences, we must also consider the validity of a VR test in differentiating between safe and less-safe drivers compared to a traditional single-screen presentation. It is possible that a VR test will better separate safe from less-safe drivers, as the unencumbered viewing position provides more opportunities for less-safe drivers to look in the wrong locations and miss the clues to the impending hazard than a 2D single-screen test can afford. Equally, however, it is possible that problems with comfort and realism in VR might reduce the ability of the test to differentiate between driver groups.

3.1.1 The current study

The findings of Study 1 suggested that our initial attempts to create a hazard test for presentation in VR headsets did not produce excessive sickness symptoms when compared with a recent meta-analysis of 55 studies (Saredakis et al., 2020). Furthermore, the hazard prediction variant of our tests was found to induce the lowest level of sickness symptoms. This result, combined with evidence from studies that argue hazard prediction testing to be at least as effective as traditional hazard perception formats in differentiating between safe and less-safe driver groups (Crundall et al., 2021; Crundall & Kroll, 2018, Ventsislavova et al., 2019), led to the prediction format being adopted for this study.

Before starting the study, we compared the quality of the clips we created for Study 1 with similar 360-degree clips created by other organisations active in the road safety field (e.g. the British Horse Society and Leicester Fire and Rescue Service). The comparison suggested that we could create better clips with a higher fidelity. Accordingly, we upgraded our 360-degree camera system and instigated a new round of clip filming and editing, building on our experience gained in Study 1.

[8] https://lissbeedesign.co.uk/project/destination-zero/

Twenty-four new clips were created and edited into both a 360-degree format and a format suitable for display on a single screen. Participants viewed 12 clips on a single screen and 12 clips within the VR headset (counterbalanced across clip sets and the order of presentation). Two participant groups were recruited: novice drivers (predominantly learners) and experienced drivers. This is a standard surrogate measure that is used for driver risk, based on the excessive crash-likelihood of inexperienced drivers (e.g. Underwood, 2007). It was predicted that the groups would differ in their accuracy at predicting imminent hazards on the single-screen test. However, whether the VR test will also differentiate between the groups is unknown. It is possible that greater distraction in the VR test may degrade performance, eroding the benefit of experience seen in the single-screen test. Alternatively, the increased immersion of the VR test may enhance the experiential benefit and lead to greater differentiation between the groups.

3.2 Method

3.2.1 Participants

Sixty-seven participants were recruited (34 experienced and 33 novice drivers). The minimum definition of an experienced driver was someone who had passed their driving test and had at least three years of active driving. Most novices were still learning to drive, though five novices had passed their driving test within the 12 months prior to the experiment. One experienced driver and six novice drivers were removed for various reasons (equipment failure, failure to give demographic information or misrepresentation of their driving status), leaving 60 valid participants.

During the study, two experienced drivers were removed owing to scores on the SSQ reaching the same threshold used in Study 1, leaving 58 participants for analysis (31 experienced drivers and 27 novice drivers). The demographics details of the participants in each of the age groups and the two participants who were removed for reasons of sickness are given in Table 3.1.

3.2.2 Design

A 2 × 2 mixed design was employed to compare driver experience (experienced versus novice drivers) across test variant (single-screen versus 360-degree hazard prediction tests). Both tests contained 12 clips from the total of 24 clips that were developed (see subsection 3.2.3). The 24 clips were split equally into two sets (set A and B) and were matched in terms of their content as closely as possible (for a full list of hazards and their descriptions, see Table 3.2). Half of the participants in each experience group saw clip set A in a 360-degree format, and clip set B in a single-screen format. The other half of the participants saw clip set A in a single-screen format, and clip set B in a 360-degree format. The presentation of the clip sets was counterbalanced across participants. Between Studies 1 and 2, Tobii released an update to the Tobii Pro Lab software, allowing the clips to be randomised within each set. The primary dependent variable for the task was the accuracy with which participants chose the correct option when predicting “What happens next?” Other dependent variables included eye-movement measures (whether or not they looked at the source of the impending hazard, their time to first fixate this source, dwell time, and fixation count) and participants’ ratings for the CRIE questions (as used in Study 1).

Table 3.1: Demographics of all participants who completed Study 2, showing participants who suffered from sickness in grey at the bottom of the table

Group	N	Gender	Mean Age (years)	Mean Driving experience (years since passing driving test)
Experienced *	31	19 females	32.8	13.2
	27	14 females	21.4	0.5
Experienced	2	2 females	57.0	36.2

Source: Authors' own (Study 2)
Note: * Not including participants who were removed owing to sickness

3.2.3 Stimuli

An Insta360 Pro 2 was used to capture new 360-degree footage at 8K resolution. The camera was mounted on the roof of a Ford Fiesta just above the driver's head to provide a similar point of view to that of the driver. Three GoPro HERO4 Silver cameras were also attached to the car to capture the view from the wing and interior mirrors. All cameras were tethered to the vehicle for safety. The footage was collected over a four-month period from July to October 2019 across Nottinghamshire, at varying times of day to capture different traffic densities. Filming was conducted for approximately one hour at a time, constrained by the battery life of the cameras. A team of traffic psychologists reviewed the footage and selected hazards on the basis of the five principles used in Study 1 (see subsection 2.2.3.1). Twenty-four clips were identified as containing suitable hazards (Table 3.2; see also Figures 5.8 to 5.11 in subsection 5.3.).

Table 3.2: Description of the 24 hazards in clip sets A and B created for Study 2

Clip number	Clip set	Paired Clip	Description	Multiple-choice options (with correct answers underlined)	Clip length (sec)
1	**A**	10	While you are approaching a left turn, the traffic lights ahead turn green. As you start to turn, some pedestrians who cross on a red man are in the middle of the road, and so you have to slow down to avoid a collision.	1. A cyclist appears from the side road on the right. 2. The pedestrians on the right step into the road. 3. As you turn left, pedestrians at the crossing step into the road. 4. An oncoming tram forces you to give way.	28
2	**A**	4	After you have waited at a set of traffic lights, they turn green and you take a left turn. As you follow the road, a white taxi does a U-turn in the middle of your lane, blocking your path.	1. A pedestrian steps into the road from the left. 2. The white car on the left performs a U-turn, blocking your path. 3. A motorcyclist overtakes you from the right. 4. A pedestrian runs across the road from the right.	23

Clip number	Clip set	Paired Clip	Description	Multiple-choice options (with correct answers underlined)	Clip length (sec)
3	**A**	22	While you are driving along a suburban route, an oncoming police car becomes visible in the distance, and so you have to pull over to give way.	1. An oncoming emergency services vehicle forces you to give way. 2. A car pulls out of the car park on the right. 3. A pedestrian steps into the road from behind the tree on the left. 4. The car ahead indicates and pulls into the side road on the right.	34
5	**A**	12	After you turn right into a wide, empty road, a pedestrian steps out from between the parked cars on the right, and crosses in front of you.	1. A pedestrian steps into the road from the left. 2. A pedestrian crosses the road from between the parked cars on the right. 3. A parked car on the right pulls off into your lane. 4. A cyclist in the cycle lane drifts into your lane ahead.	44
6	**A**	16	While you are driving along a busy road on a campus, a chain of pedestrians obscured by parked vehicles enter the road, forcing you to stop.	1. A worker steps out from behind the van on the left. 2. The van parked on the right pulls off in front of you. 3. A cyclist appears from behind the van parked on the right. 4. Pedestrians step into the road from the left ahead.	44
7	**A**	15	As you are travelling along a busy urban route, an oncoming car pulls across your path into the side road on the left, after a set of traffic lights.	1. The van parked on the left pulls off in front of you. 2. A pedestrian steps into the road from the left at the crossing ahead. 3. A pedestrian steps into the road from the left. 4. An oncoming car pulls across your path into a side road on the left.	34
8	**A**	23	While you are driving along a main arterial road with a cycle lane, a delivery van parked on the kerb behind some trees pulls off in front of you.	1. The van ahead pulls off in front of you. 2. A car pulls out of the car park on the left. 3. A pedestrian steps into the road from the left. 4. A pedestrian steps into the road from the right.	24
9	**A**	20	While you are following a delivery van, a car can be seen ahead obstructing the path, causing the van and you to brake.	1. The door of the parked car on the left opens. 2. A worker steps into the road from behind the barriers on the right. 3. A car turning in the road ahead forces you to brake to avoid a collision. 4. The car parked on the forecourt on the left starts to reverse into the road.	39

Clip number	Clip set	Paired Clip	Description	Multiple-choice options (with correct answers underlined)	Clip length (sec)
11	**A**	24	While you are driving in congested traffic, a bus on your left indicates to pull into your lane to overtake a cyclist.	1. The bus on your left pulls into your lane. 2. A car undertakes you on the left at speed. 3. A cyclist pulls into your lane from the left. 4. A car pulls out from the side road on the right.	26
13	**A**	14	As you are driving along a quiet residential road, with parked vehicles along either side, a car suddenly appears from behind a parked car	1. A pedestrian steps out from behind the parked van on the left. 2. A car emerges from behind a parked car on the left. 3. A parked car on the right pulls off into your lane. 4. An oncoming car encroaches on your lane, forcing you to give way.	30
17	**A**	21	While you are travelling along a main road with a bus lane, a pedestrian begins to cross from the island in the middle.	1. The parked car on the left pulls off in front of you. 2. A car pulls out of the side road on the left across your path. 3. A pedestrian steps into the road from behind the parked car on the left. 4. A pedestrian steps into the road at the island ahead.	25
18	**A**	19	While you are driving along a back street with parked cars, an oncoming car moves into the middle of the road to overtake the parked cars, forcing you to slow down and give way.	1. A van appears from the side road on the left. 2. A squirrel runs across the road from the hedges on the left. 3. An oncoming car encroaches on your lane, forcing you to give way. 4. The door of a parked car on the right opens and a pedestrian steps out.	15
4	**B**	2	As you are driving along a suburban road with parked cars and speed bumps, a car starts to pull out, blocking your path.	1. A door of the parked car on the left opens. 2. A pedestrian steps out from behind the parked car on the right. 3. A car on the left pulls out into the road, blocking your path. 4. A car pulls out of the driveway of the house on the left.	40
10	**B**	1	While you are turning left into a side road amongst some roadworks, a pedestrian obscured by the van ahead of you becomes visible, causing you to slow.	1. The construction vehicle on the right swings into your lane as it turns. 2. The white van that turned into the road has to brake suddenly. 3. A worker steps into the road on the right from behind the barriers. 4. A pedestrian in the road forces you to slow down as you turn left.	34

Clip number	Clip set	Paired Clip	Description	Multiple-choice options (with correct answers underlined)	Clip length (sec)
12	**B**	5	As you are pulling out of a road and turning right to join a main road, a pedestrian steps out and crosses the road in front of you.	1. The car ahead reverses towards you, as it is over the give way line. 2. A pedestrian steps in front of you from the left. 3. A pedestrian steps in front of you from the right. 4. A car turning into your road encroaches on your lane.	30
14	**B**	13	While you are turning into a residential road with parked cars, a car suddenly appears from a side road on the right.	1. A car emerges from behind the parked car from the side road on the right. 2. A pedestrian steps out from behind the parked car on the right. 3. A parked car on the left pulls off into your lane. 4. An oncoming car encroaches on your lane, forcing you to give way.	30
15	**B**	7	While you are driving along a main dual-lane road, an oncoming car pulls across your path into a side road on the right.	1. A car pulls out of the side road on the left into your path. 2. The van ahead brakes suddenly because of congestion ahead. 3. An oncoming car pulls across your path into a side road on the left. 4. A pedestrian steps into the road at the crossing ahead.	28
16	**B**	6	While you are driving through a busy city-centre road, a pedestrian steps out into the road from a bus stop on the left.	1. The oncoming car performs a U-turn in front of you. 2. A pedestrian steps into the road from the left. 3. Pedestrians step into the road from the right. 4. An oncoming emergency services vehicle forces you to give way.	25
19	**B**	18	While you are passing through a town centre, a bus ahead pulls over, forcing an oncoming car to encroach on your lane.	1. An oncoming car encroaches on your lane, forcing you to give way. 2. A car pulls out of the side road on the left. 3. A parked car on the right performs a U-turn and blocks your path. 4. A pedestrian runs across the road from the right to reach the bus stop.	26
20	**B**	9	On a busy city-centre road, a bus pulls out in front of you from a bus stop on the left.	1. The pedestrian on the left steps into the road. 2. The bus on the left pulls into your lane. 3. A pedestrian steps off the central island into your lane. 4. An oncoming car pulls across your path into a side road on the left.	24

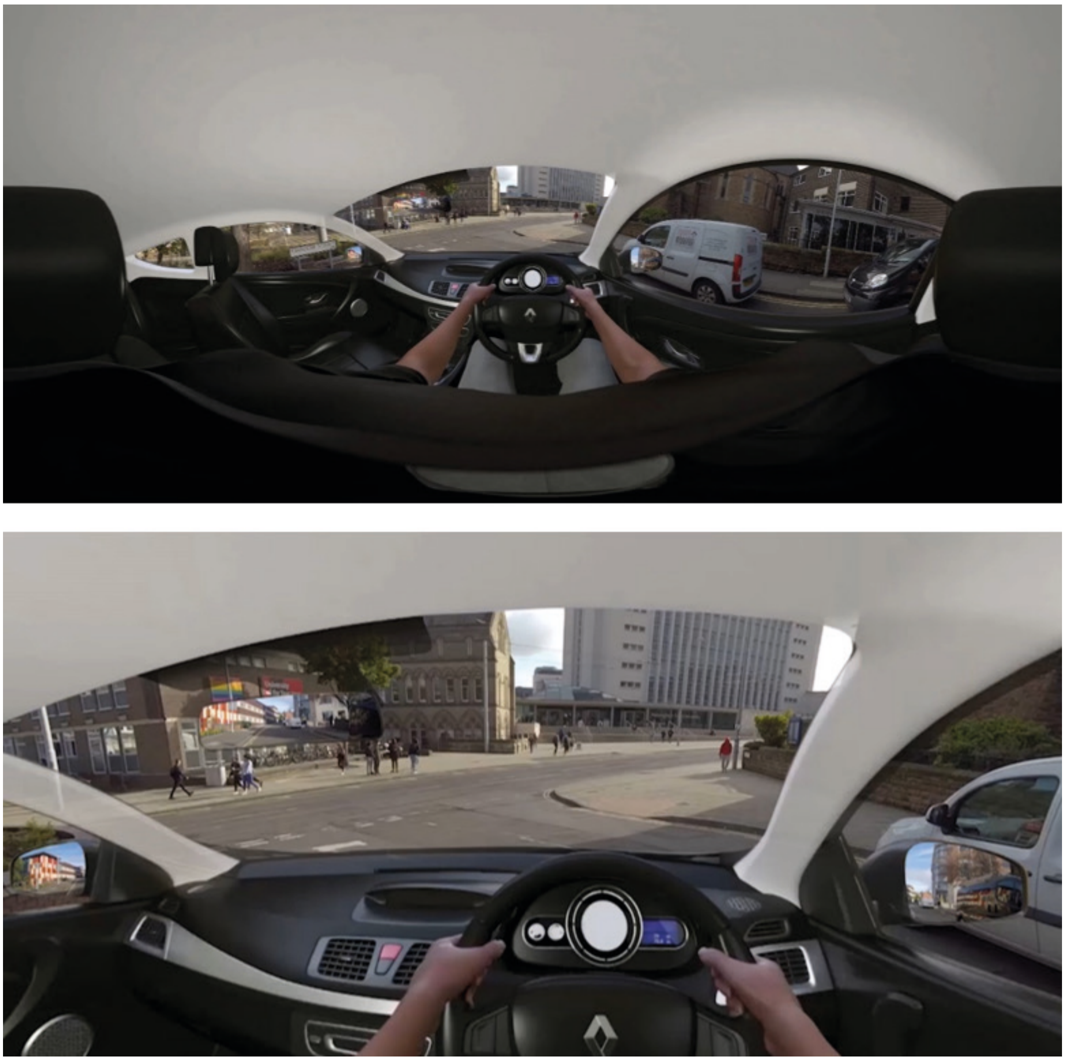
Clip number	Clip set	Paired Clip	Description	Multiple-choice options (with correct answers underlined)	Clip length (sec)
21	**B**	17	While you are driving along a busy arterial road with parked cars either side, a pedestrian runs out across your path from the right.	1. The passenger door of the van parked on the right opens. 2. A worker carrying scaffolding steps into the road from the left. 3. A car pulls out from the side road on the left. 4. A pedestrian crosses the road from between the parked cars on the right.	25
22	**B**	3	While you are travelling along a busy urban route, an oncoming van encroaches on your lane as it overtakes a parked vehicle.	1. A car pulls out of the side road on the left into your path. 2. A pedestrian steps out from behind the parked car on the left. 3. A car parked on the right performs a U-turn in front of you. 4. An oncoming van encroaches into your lane.	30
23	**B**	8	On a busy main road with a bus lane, as you approach a set of traffic lights, a van appears from a side road on the left.	1. A pedestrian steps into the road from the bus stop on the left. 2. A van pulls out of the side road on the left. 3. A pedestrian steps off the central island into your lane. 4. An emergency service vehicle undertakes you at speed.	25
24	**B**	11	On a main suburban road with a bus lane, a bus indicates to pull into your lane to avoid a bus pulling up on the left.	1. The bus on your left pulls into your lane. 2. A pedestrian steps into the road from behind the tree on the left. 3. The car in the right lane cuts in front of you. 4. A cyclist at the island enters into the road from the right.	31

Source: Authors' own (Study 2)

A new graphic overlay of a car interior was developed, including the body and arms of a virtual driver. Footage from the cameras was synchronised and edited into the graphic overlay to produce the final version of the hazard prediction clips.

As with Study 1, the clips were silent apart from the voice-over of an 'instructor' giving directions (e.g. "take the next left"). These voice-over instructions were included to ensure that the viewer shared the intentions of the film car driver and would therefore make more appropriate eye movements for the intended path of the vehicle (see Crundall et al., 2021). The equivalent single-screen clips were cropped from the 360-degree clips, restricting the visual field to the outside edge of one side mirror across to the other mirror (constituting approximately 31% of the full 360-degree view; see Figure 3.1). This is the standard area of the visual field that is typical in single-screen hazard tests (e.g. Ventsislavova et al., 2019). The visual content of the two versions were identical, as they were taken from the same footage, the only difference being the presentation modality. The average overall clip length (this being also the time up to the occlusion point) across both clip sets was 30 seconds (range: 15 s to 44 s).

Figure 3.1: Screenshot of the full 360-degree view of the video-based test created for Study 2 (top panel) and the cropped area used in the single-screen variant (bottom panel)



Source: Authors' own (Study 2)

Additional stimuli included the same questionnaires used in Study 1 (demographics, SSQ and CRIE), and the Road Safety Scotland practice clip. The full version of the practice clip was used (2 minutes and 53 seconds).

3.2.4 Apparatus

The single-screen test was presented on a computer monitor measuring 48.3 cm × 30.5 cm. The full screen clips subtended a visual angle of 44 degrees by 28.5 degrees when participants were seated at a distance of 60 cm. The monitor was connected to a SensoMotoric Instruments' remote eye-tracking device (SMI RED500), sampling at 500 Hz. Participants were provided with a mouse to select one of the four multiple-choice options following clip occlusion. The directional voice-over was played through speakers.

As in Study 1, the 360-degree test was presented to participants in an HTC Vive headset with an integrated Tobii Pro eye tracker (2,160 × 1,080 resolution, 120 Hz). Again, a Republic of Gamers ASUS ROG Strix Hero III Gaming Laptop was used to administer the tests using Tobii Pro Lab software to design and present the experiments. Participants listened to the directional voice-over via headphones integrated with the VR headset.

3.2.5 Procedure

Participants were invited to the laboratory at Nottingham Trent University. Upon arrival, participants were given instructions and asked to sign a consent form, which detailed their right to withdraw at any point without explanation, and to withdraw their data from the study at a later point. They then completed the demographics questionnaire. Participants completed both the single-screen and 360-degree version of the test in a counterbalanced order.

For the single-screen test, participants were seated approximately 60 cm from the screen. 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 a hazard begins to occur. Following this, they were presented with four possible options regarding what might happen next, from which they had to select the correct answer using the computer mouse.

For the 360-degree test, 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 presented in 360 degrees, taken from the perspective of a driver. They were informed that the clip would suddenly stop just as a hazard begins to materialise. Following this they were presented with four possible options (numbered 1–4) regarding what might happen next, from which they had to select the correct answer. Once they had chosen the correct answer, they verbally gave the researcher their chosen option. The testing session lasted around 30–45 minutes. The simulator sickness checklist was administered at three different time points: prior to the experiment (time point 1), following the practice clip, and at the end of the VR test. See subsection 2.2.3.3 for criteria for participant withdrawal owing to sickness. At the end of each test, participants completed the CRIE questions for the test they had just completed. All participants received a £10 Amazon voucher for taking part in the study.

3.3 Results

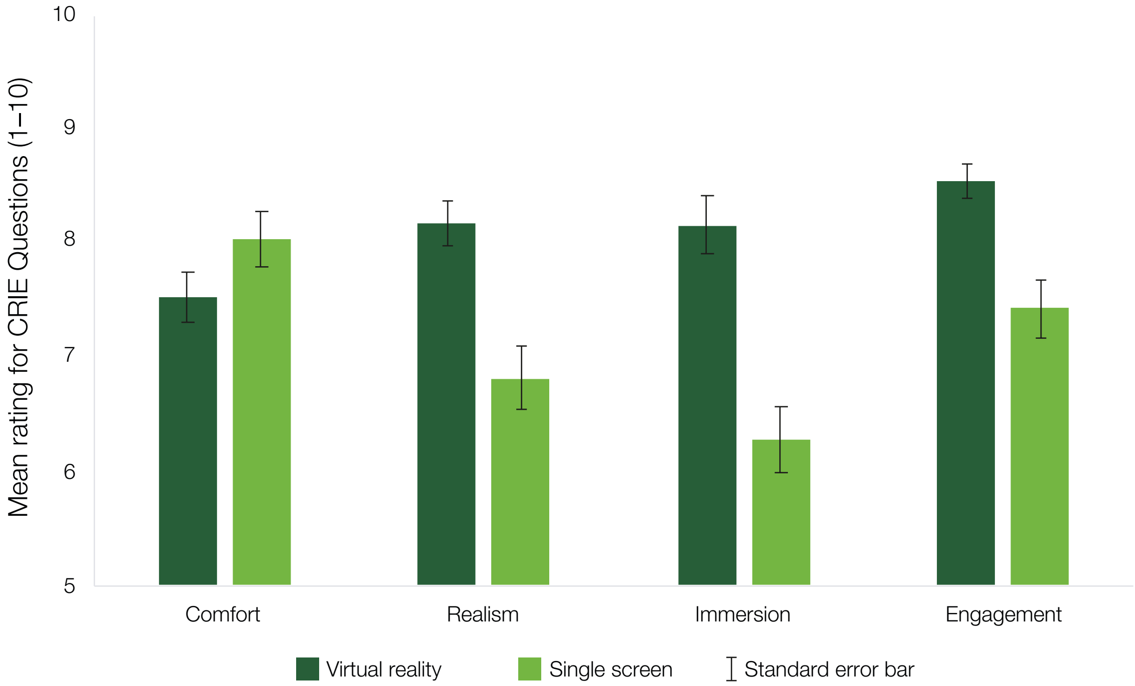
3.3.1 Comfort, realism, immersion and engagement questions

Participants gave ratings for each of the questions regarding CRIE for both the 360-degree and single-screen hazard prediction tests. In addition to the two participants removed for sickness (subsection 3.2.1), one further participant was excluded owing to missing CRIE data (though their data was included in the behavioural analysis).

Participants' ratings for each measure were entered into a series of 2 × 2 mixed ANOVAs across driver group (experienced vs novice), and presentation mode (VR vs single screen).

Ratings for realism, immersion and engagement all produced main effects. Participants rated the 360-degree clips as more realistic than the single-screen clips (8.2 vs 6.8; $F(1, 55) = 24.4$, $MSE = 2.2$, $p < .001$, $\eta^2 = .31$). They also believed that the 360-degree clips were more immersive (8.2 vs 6.3; $F(1, 55) = 26.6$, $MSE = 3.8$, $p < .001$, $\eta^2 = .33$), and more engaging (8.5 vs 7.4; $F(1, 55) = 17.1$, $MSE = 2.1$, $p < .001$, $\eta^2 = .24$). No other main effects or interactions approached significance (all values of $p > .05$). Figure 3.2 shows the group means.

Figure 3.2: Average ratings given for each of the four items on the Study 2 CRIE questionnaire for the virtual reality and single-screen tests



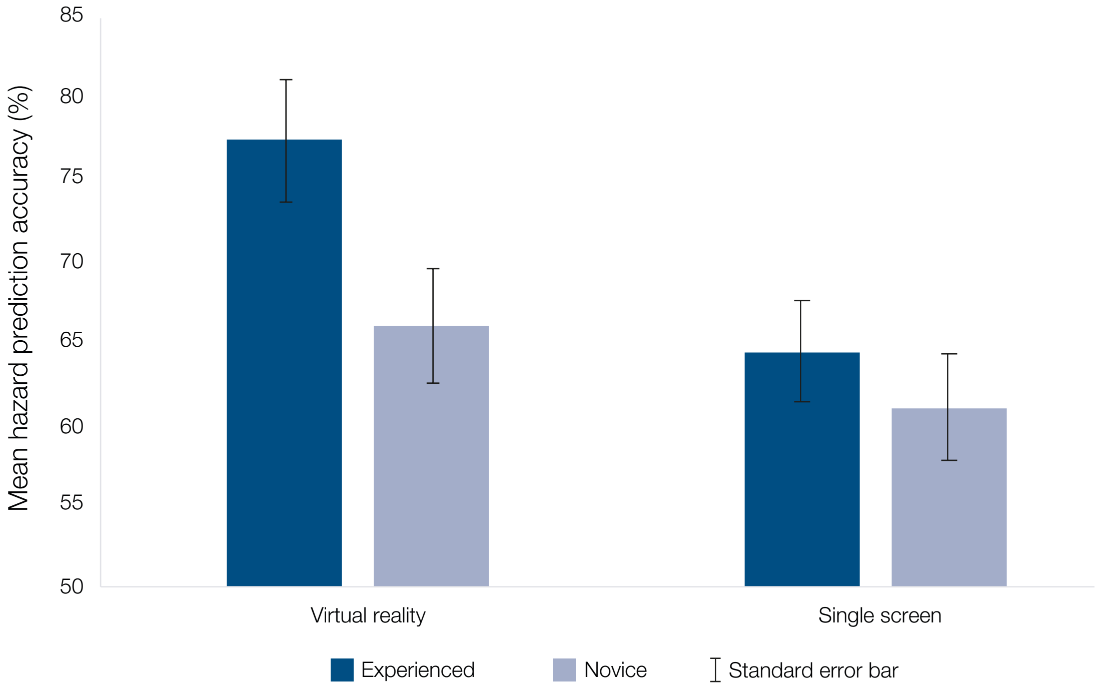
Source: Authors' own (Study 2)
Note: * p < .001

3.3.2 Hazard prediction performance

Each participant saw all 24 clips, half in VR and half on the single screen. The two clip sets were counterbalanced across participants in both their order of presentation and presentation mode. Three clips evoked particularly poor performance from all participants, suggesting that these three hazards were too difficult to spot (clips 9, 12 and 17). Only 21% of participants identified the correct response on these clips, which was below mean chance expectancy. All data referring to these three clips was therefore removed from the subsequent analysis.

Participants' percentage scores on each test were subjected to a 2 × 2 mixed ANOVA (group experience × presentation mode). This analysis revealed a main effect of presentation mode, $F(1, 56) = 7.0$, $MSE = 333.9$, $p = .01$, $\eta^2 = .11$, demonstrating that drivers predicted significantly more hazards when viewing the clips in VR headsets than when viewing the clips on the single screen (72.2% vs 62.9%). There was also a main effect of experience, $F(1, 56) = 4.7$, $MSE = 338.4$, $p = .04$, $\eta^2 = .08$, demonstrating that experienced drivers predicted significantly more hazards than novice drivers (71.0% vs 63.6%). Though Figure 3.3 suggests that the VR condition is responsible for the main effect of driving experience, the interaction does not reach the threshold for significance ($F(1, 56) = 1.4$, $MSE = 333.9$, $p = .25$, $\eta^2 = .02$). Nonetheless, the pre-planned comparisons of the driver groups for each presentation mode show that experienced driver and novice driver performance on the 360-degree clips differs significantly (77.4% vs 66.1%, $p = .03$), while the difference noted in the single-screen condition does not reach the threshold of significance.

Figure 3.3: Hazard prediction performance for the virtual reality and single-screen tests of Study 2

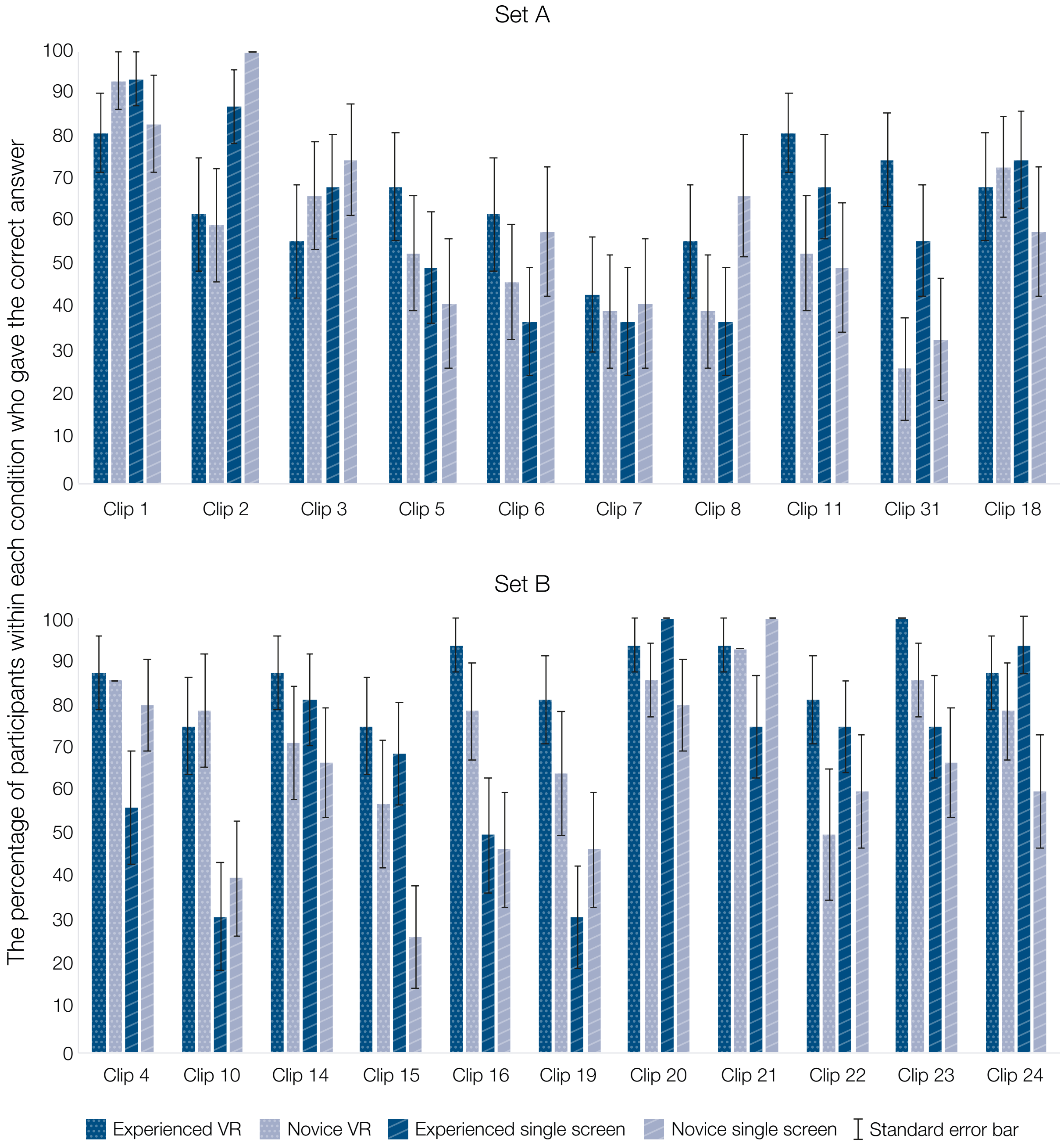


Source: Authors' own (Study 2)

To assess the contribution of individual clips to these effects, the accuracy for each hazard for each presentation type (VR and single-screen) was also charted. Figure 3.4 provides the percentage of participants that chose the correct option for each clip (separated into the two clip sets). As can be seen, experienced participants were more accurate overall at predicting the hazards than novices in both presentation types.

It is clear from this figure that the clips differ in their pattern of correct responses across the conditions. Several clips show that the VR condition is associated with an ostensibly larger difference between experienced drivers and novice drivers (see for example clips 5, 11, 13, 16 and 22). Interestingly, there are several clips where the apparent superiority of the 360-degree clips displays a reversal of the expected trend, with novices outperforming the experienced drivers in the single-screen condition (e.g. clips 4, 6, 8, 19 and 21). In these clips, it appears that novices receive some benefit from the single-screen condition that is not commensurate with the crash risk of their group (perhaps some form of attentional funnelling that the experienced drivers resist). When placed into a VR headset, however (which is more akin to the real world), the novice advantage on these clips disappears. There are, however, a handful of clips where the VR presentation mode has, far from demonstrating the ability to differentiate between novice and experienced drivers, actually yielded the opposite outcome (e.g. clips 15, 18, 20 and 24).

Figure 3.4: Hazard prediction accuracy scores across all clips (clip sets A and B) for the virtual reality and single-screen tests of Study 2



Source: Authors' own (Study 2)

Note: Clips 9, 12 and 17 were removed after being adjudged too difficult to spot.

3.3.3 Eye-movement measures

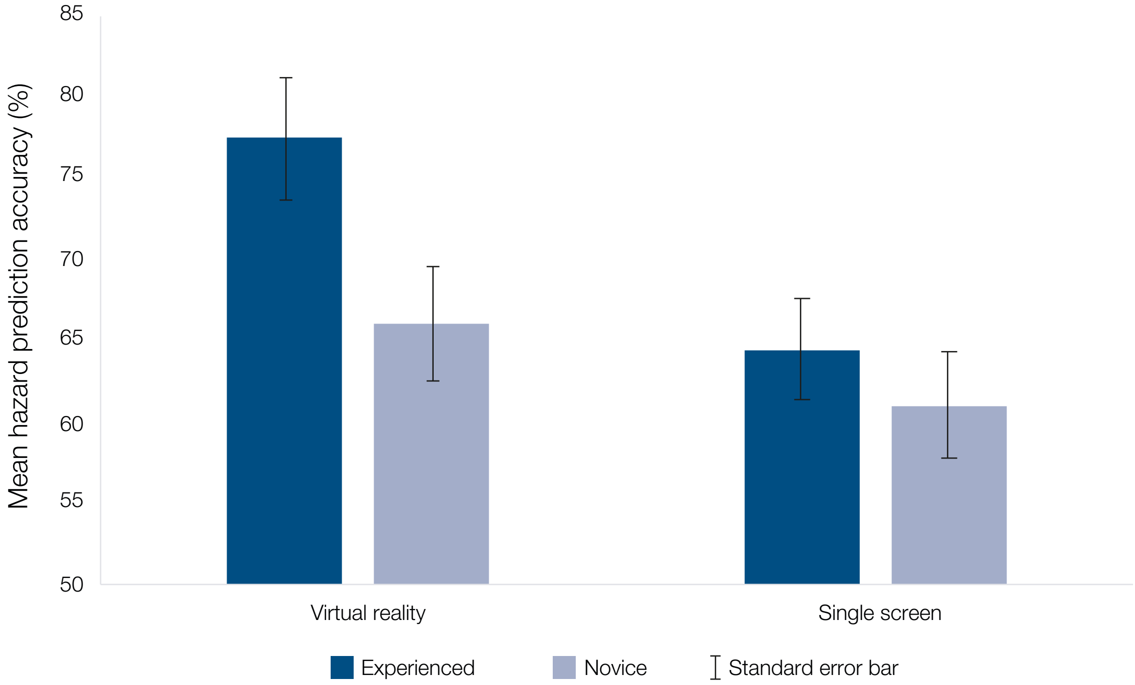
A number of eye-movement measures were calculated that reflected whether participants looked at precursors to each hazard (up to the point of occlusion), and if so, how quickly they looked at them, and for how long. Fixations on hazards were defined by creating temporal and spatial areas of interest (AOIs) for each video clip in the respective eye-tracking software used for both eye-tracking systems used in this study. These AOIs lasted for the duration of the hazard precursor window, and accepted only fixations that fell on the hazard (+1 degree of visual angle, approximately). For instance, in a clip when an oncoming car will eventually turn across the viewer's path, the precursor window starts when the oncoming car is first visible and ends at the point of occlusion as it begins to turn across their path. The size and positioning of the AOIs in both eye-tracking systems were matched by using visual landmarks in the videos to identify the same locations in both systems. Any fixations of shorter duration than 60 ms were discarded. Across all analyses of eye data, five participants (all novices) were removed owing to poor calibration in the single-screen test, and a further participant (experienced) was removed for poor calibration in the VR headset.

3.3.3.1 Did they spot the hazards?

The most basic question that eye-movement analyses can address is whether participants look at crucial elements of the scene. In the current study, we were concerned with whether the drivers looked at the hazardous precursors prior to occlusion. A precursor precedes a hazard and acts as a clue to the upcoming hazard. For instance, a pedestrian on the pavement walking towards the road may lead to the prediction that the same person may step out into the road and become a hazard.

The percentage of hazardous precursors that drivers (both novice and experienced) fixated was calculated (out of 10 for clip set A and 11 for clip set B). However, a further 12 participants in the single-screen condition showed poor calibration on an average of 1.75 clips. For these participants, the percentage of hazard precursors that they fixated was calculated out of the total number of hazard windows for which there was sufficient eye data. This data was subjected to a 2 × 2 ANOVA between-groups across driver group and presentation mode. Though Figure 3.5 suggests that experienced drivers might have looked at more hazard precursors than novices in the VR headset, the analysis did not reveal any significant effects.

Figure 3.5: Average percentages of hazards that participants looked at in the virtual reality and single-screen tests of Study 2

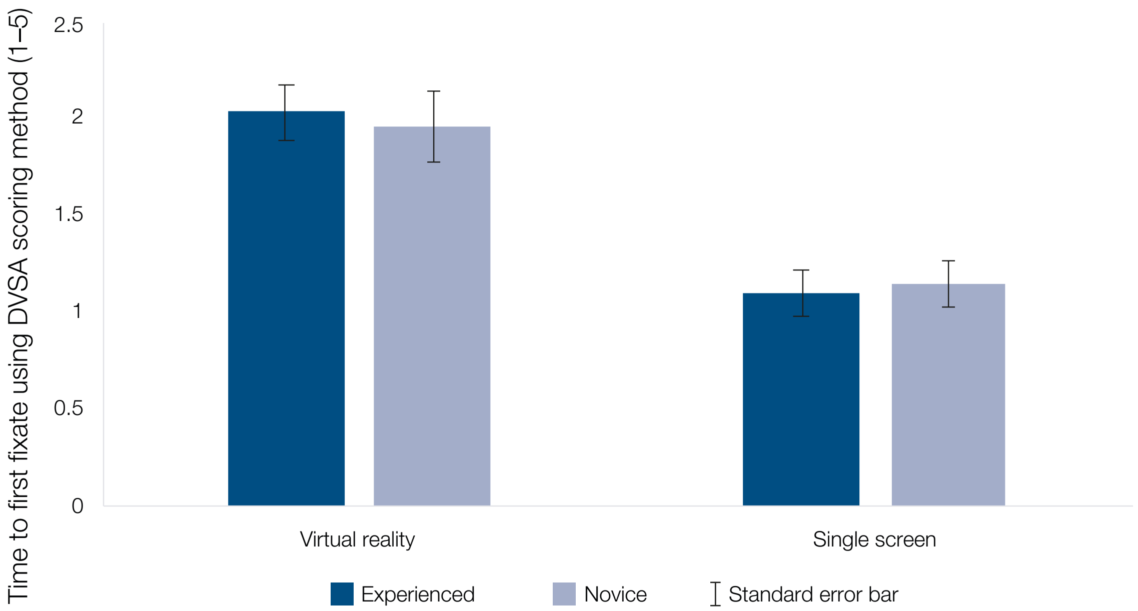


Source: Authors' own (Study 2)

3.3.3.2 How soon do participants fixate the hazard precursors?

The **time taken to first fixate** each hazard precursor was calculated as the time at which participants first looked at the precursor minus the time at which the precursor was first visible. This is a measure of how quickly participants spot the hazard precursor in each clip. The method used by the DVSA to score the national hazard perception test was applied to our time-to-first-fixate measure: the precursor windows for each hazard were split into five even sections, with five points awarded for a fixation in the AOI in the first section, four points for a fixation in the AOI in the second section, and so on. The data was compared across driver experience and presentation type using a 2 × 2 mixed ANOVA. This revealed a main effect of presentation mode, $F(1, 50) = 48.7$, $MSE = .4$, $p < .001$, $\eta_p^2 = .49$, with those participants who saw the test in the VR headsets fixating precursors faster than those who saw them on a single screen (2.26 points vs 1.12 points). Experience did not produce an effect, and the interaction was not significant (Figure 3.6).

Figure 3.6: Average number of points scored, reflecting how fast participants fixated the hazardous precursor in each test type in Study 2



Source: Authors' own (Study 2)

3.3.3.3 Amount of attention devoted to the hazard precursors

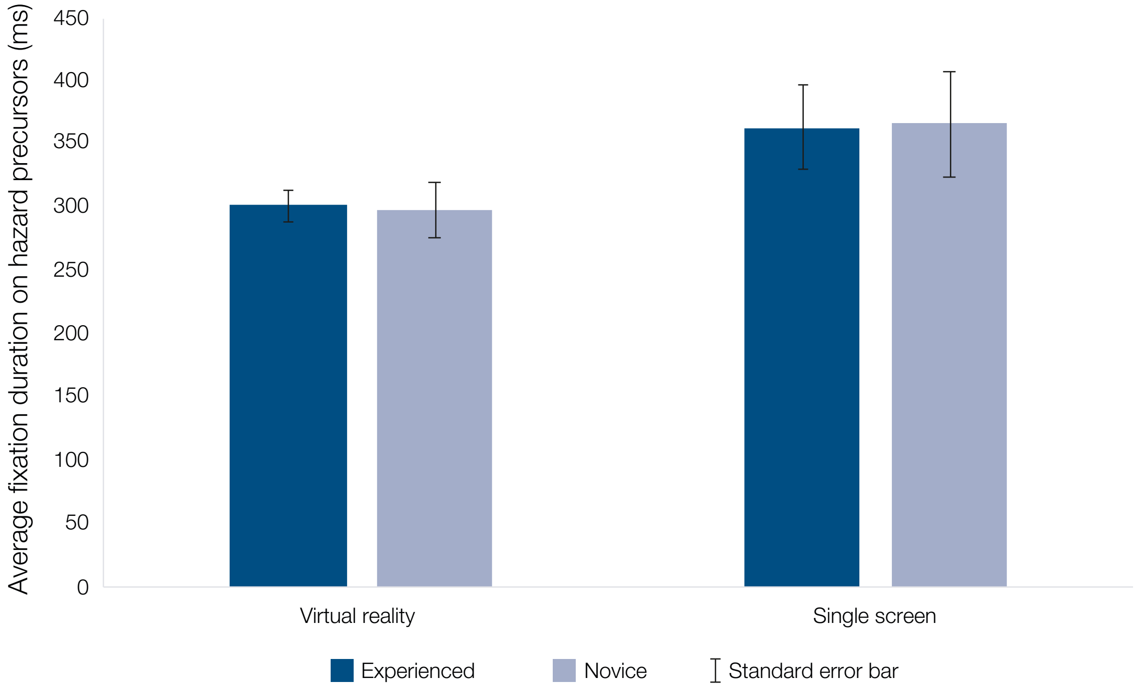
The amount of time that each participant devoted to the hazardous precursors was also calculated. Measures of attention to these precursors reflect the preparatory work that drivers undertake in actively predicting imminent hazards. For the current analyses, the measures of mean fixation duration, first-fixation duration and dwell time were chosen to reflect attention given to the hazard precursors.

Comparison of **mean fixation durations** on precursors revealed a main effect of test type ($F(1, 50) = 5.8$, $MSE = 31{,}712.5$, $p = .02$, $\eta_p^2 = .10$), with single-screen precursors receiving significantly longer mean fixation durations than precursors in the VR headsets (365 ms vs 300 ms, respectively). Driver group did not produce an effect, and the interaction was not significant (Figure 3.7).

A similar analysis of the mean **duration of** participants' **first fixations** on the precursors also revealed a main effect of test type ($F(1, 50) = 6.3$, $MSE = 22{,}481.1$, $p = .02$, $\eta_p^2 = .12$), with participants having longer first-fixation durations in the single-screen test than the VR test (363 ms vs 289 ms). Driver group did not produce an effect, and the interaction was not significant (Figure 3.8).

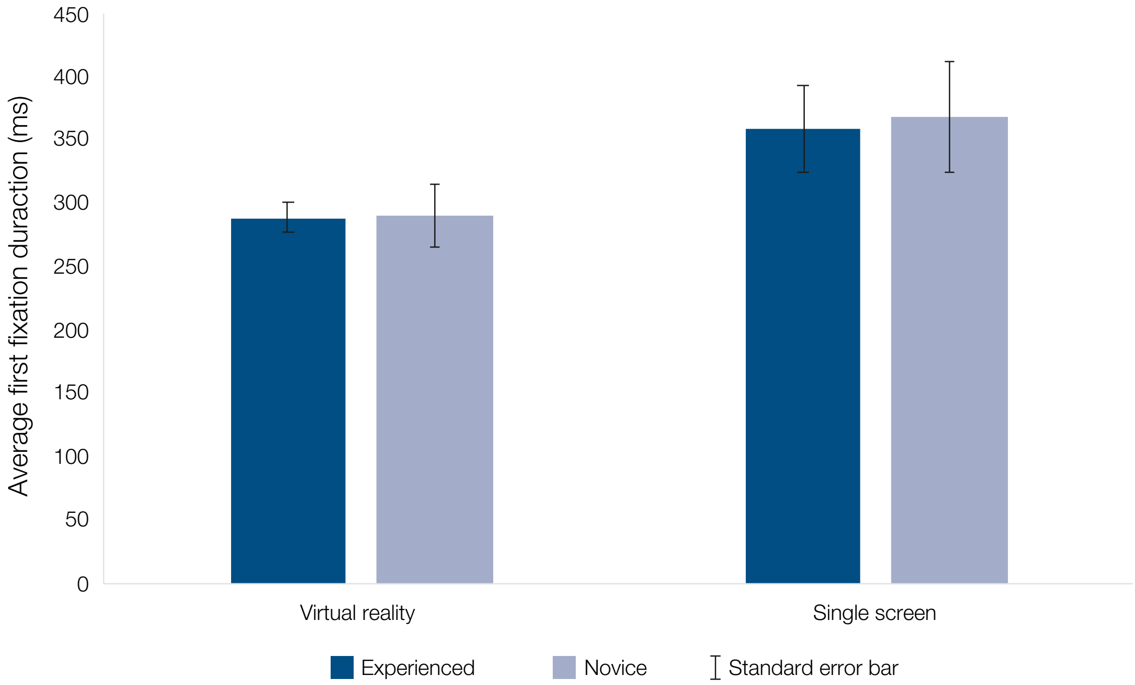
The **dwell-time** measure was calculated as the percentage of time the participants spent looking within the hazard precursor window (as a function of how long it was available for inspection). Once again we found a significant effect of test type ($F(1, 50) = 11.4$, $MSE = .02$, $p = .001$, $\eta_p^2 = .19$), with those who saw the test in the VR headsets having spent longer looking at the hazardous precursors than those participants who viewed the test on the single screen (26.0% vs 16.2%, respectively; see Figure 3.9 and the footnote to the discussion section which follows), but the experiential factor and the interaction were not significant.

Figure 3.7: Average fixation duration for each group on the hazardous precursors for each test type in Study 2



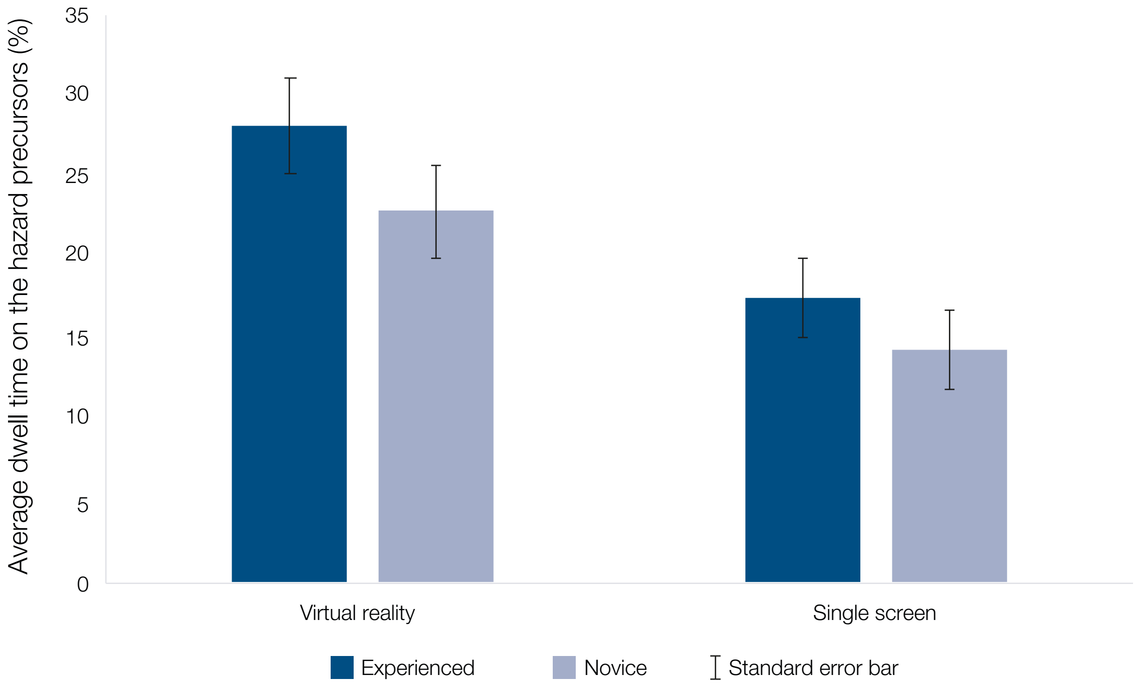
Source: Authors' own (Study 2)

Figure 3.8: Average first-fixation duration for each group on the hazardous precursors in Study 2



Source: Authors' own (Study 2)

Figure 3.9: Average dwell time (%) on the hazard precursors in Study 2 across the different participant groups



Source: Authors' own (Study 2)

3.4 Discussion

The primary aim of Study 2 was to compare participant experiences, and behavioural and oculomotor data resulting from a hazard prediction test, across two presentation modes (360-degree vs single-screen). Specifically, we were interested in whether the participants reported the VR test to be more or less comfortable, realistic, immersive and engaging than the single-screen test. Regarding the behavioural data, we wanted to ascertain whether the VR test was more (or less) effective at differentiating between driver groups on the basis of a surrogate risk measure of driving experience.

The results of the CRIE analysis suggested that drivers did not perceive the VR version of the test to be significantly less comfortable than viewing the test on a single screen. While we expected the VR headset to have a negative impact on overall comfort, and the mean ratings did tend towards that direction, the difference did not breach the threshold for statistical significance. What is more, withdrawal of participants owing to sickness was even lower than the number removed in the previous study (3% in Study 2, compared to 5% in Study 1).

The other three CRIE measures did, however, produce the predicted results, with the VR presentation mode being reported as providing a more realistic, immersive and engaging experience. The mean ratings of these three measures for the VR condition closely mirror those found in Study 1, with ratings for realism and immersion hovering around 8 out of 10, rising to approximately 8.5 for engagement in both studies.

The behavioural data suggests that all drivers found it easier to predict the hazards in the 360-degree clips. While it might be tempting to attribute this to greater reported immersion and engagement, it is also likely that the increased viewing angle that our VR hazards subtended on the retina also played a role. The amount of information presented in the single-screen condition (Figure 3.1) was selected to emulate other hazard perception and prediction tests that have previously included both side mirrors (e.g. Ventsislavova et al., 2019). This inevitably means that the visual angle that the hazard subtends on the retinas in the single-screen condition is less than the visual angle produced by the same hazard in the VR headset. The single-screen view approximates to 31% of the 360-degree field (i.e. 112 degrees), but it is, by necessity, presented on a single screen with a visual angle of 44 degrees in the horizontal plane. This means that objects appearing in the VR headset are approximately 2.5 times the size of the same object in the single-screen condition. This fact is no doubt the primary cause of all the effects noted in the eye-movement data, which confirm that the hazards in the VR presentation mode were easier to see: drivers spotted VR hazard precursors sooner and more often than single-screen precursors, whereas the single-screen precursors evoked longer fixation durations, as drivers found it harder to process these visually smaller threats.[9]

This, however, does not imply that we can simply declare VR to be ‘better’ because drivers spot more hazards when using it. It merely means that comparable hazards are easier to see in VR headsets because they appear larger in the visual field. It would have been possible to better equate the visual angle of hazards in the single-screen condition to that of the VR condition, but this would have made a poor hazard test for presentation on a single screen. The amount of the visual world made visible would have had to be restricted to such a narrow cone of vision that it would have been impossible to spot any hazards that originated even slightly off the line of direct heading, as they would have fallen outside the screen’s visible area. To create a valid single-screen test, we must present drivers with as much of the scene as possible, to enable them to detect precursors as they develop (e.g. a car in a side street approaching the road being driven on).

While we cannot therefore conclude that higher overall accuracy rates in the VR condition mean that the 360-degree hazard clips are better than the single-screen comparators, this difference between the conditions does provide us with one indirect benefit of VR presentation: as hazards appear larger in a VR test compared to a single-screen test, we can use precursors that are further away from the film vehicle. When creating a single-screen hazard test, one cannot expect drivers to read the road at the most extreme distances shown on screen, as the resolution is not sufficient for them to make use of any information they gather. In VR, however, future tests can employ hazards and precursors that initially appear at much further distances than in a single-screen test. This will encourage trainees to look further down the road when trying to anticipate hazards, which is one of the key behavioural distinctions marking out the most highly trained drivers (e.g. Lappi, Rinkkala & Pekkanen, 2017).

[9] The analysis of dwell time reveals the opposite effect to that noted in the analyses of fixation durations: dwell time on precursors is higher for 360-degree clips than for the single-screen condition, despite fixation durations being shorter. This suggests that viewers can process the precursor faster in VR and then decide to return to look at it depending on its likelihood of becoming a hazard. Thus, while individual fixations might be shorter on 360-degree precursors, participants probably fixate the precursor more often, thereby increasing their total dwell time.

More important, however, than a simple increase in accuracy in the VR test, is how the use of 360-degree clips influences the gap in accuracy rates between novice and experienced drivers. As noted in the introduction, it was possible that the use of 360-degree clips could reduce – or increase – any gap between the two driver groups. For instance, the greater ease with which all drivers could spot hazards in the VR headset could have resulted in a ceiling effect, with both groups performing equally well, reducing the gap. Alternatively, the ability to look wherever one wants to in the 360-degree clips may provide novices with greater opportunities to look in the wrong locations, thus increasing the performance gap between our groups.

The results revealed a main effect of group, with experienced drivers outperforming novice drivers regardless of the test, but the interaction between the two factors (mode and group) was not significant. Pre-planned comparisons, however, suggest that the groups differ in performance only in the VR condition, offering weak but promising evidence for VR superiority. When one looks at performance on the individual clips, the equivocal nature of the data becomes clear. Several clips show a clear advantage in using VR to separate out the two driver groups, though several more clips show no ostensible trend. Furthermore, a handful of clips suggest that the transition to VR can actually degrade their ability to differentiate between novice and experienced drivers. While inferential statistics were not undertaken on individual clips (as family-wise error would render such analyses insensitive), the pattern of results clearly suggests that some clips benefit from presentation in VR, while others do not. This heterogeneous pattern is not without precedent (Crundall et al., 2021), and is useful in providing pointers for future iteration, as we refine the test to improve its efficacy.

In conclusion, Study 2 has shown great promise for VR-based hazard prediction assessment. Participants clearly rate the VR experience as more realistic, immersive and engaging, yet the reported additional discomfort of wearing a VR headset was not sufficient to breach the statistical threshold for significance. Regarding the behavioural data, the VR test was certainly not worse than the single-screen condition in differentiating between the driver groups, and pre-planned contrasts offer weak evidence that it may even be superior to the single-screen test. While it is certainly true that this conclusion does not hold for every clip, it can be said, on the basis of this first pass at a VR hazard prediction test, that there appears to be much to recommend VR as a presentation mode, and no reason to warn against it.

4. Study 3: A Comparison of a Computer-Generated Imagery 360-Degree Hazard Test and a Single-Screen Hazard Test

4.1 Introduction

Study 2 demonstrated self-reported data in favour of a 360-degree hazard prediction test over a single-screen one, and weak behavioural data suggesting that the VR version was better also at differentiating between driver groups, at least within a subset of video-based clips. However, while video-based hazard tests predominate in the research literature, users of hazard perception assessments tend to favour clips that are created with CGI. The DVSA moved from using video clips to CGI in 2015, and the Dutch agency charged with updating their national hazard test has opted for CGI. Several driver training companies also prefer CGI clips.

The argument for CGI clips is based primarily on the need to periodically update clips to reflect current road rules and new car models. While such changes mean that video clips need to be refilmed, CGI clips can merely be re-rendered with newer car models replacing the older versions, and other needed updates.

The disconnect between the research field (using predominantly video-based clips) and stakeholders (using CGI) is possibly explained by the initial cost and/or expertise required to develop CGI clips. Despite some researchers creating CGI hazard tests (e.g. Malone & Brünken, 2016; Crundall et al., 2021), the majority of research groups use video-based recordings because the initial cost is cheaper (especially when one is merely recording clips for a research study, rather than for a national test). The use of CGI also means that the hazards must be staged (i.e. programmed), and while the original video-based UK hazard test also included staged scenarios, many researchers now argue in favour of naturalistic hazards (see Moran et al., 2019).

Thus, we are currently in the interesting situation where major stakeholders are investing in CGI hazard tests, yet the research evidence for using animated clips is sparse. Furthermore, any decision to use CGI clips necessitates that these hazards are designed by experts, rather than simply showing hazards that occur naturally on real roads. If one assumes that experts can specify all the relevant nuances of hazards that would allow safe drivers to detect these dangers in the real world, and, moreover, that the programmers can translate these nuances into a clip with sufficient complexity and fidelity, then CGI clips should, in theory, be as effective as video clips. Certainly, recent research (Crundall et al., 2021) demonstrated strong evidence that CGI clips were able to differentiate between safe and less-safe drivers, though some clips were found to be more successful than others. Notably, one of the less-successful clips in the Crundall et al. (2021) study was the appearance of a bicycle on the near side of the vehicle. The authors speculated that the restricted view of the single-screen presentation mode may have restricted drivers' ability to spot this hazard. This particular hazard may have fared better if it had been presented in 360 degrees.

For VR presentation, CGI clips may offer some additional advantages. For instance, live recording from a moving vehicle often results in image wobble in the footage. Even with image stabilisation software and post-production editing, it is difficult to provide a video clip that is devoid of jiggle. CGI clips, however, can provide a perfectly smooth viewing perspective, which may improve feelings of comfort in the VR headset. A second advantage is that clips can be programmed to take advantage of the 360-degree presentation, using hazards that might not normally be spotted on a restricted single screen (such as the appearance of the near-side cyclist in Crundall et al., 2021).

Despite the theoretical benefits of CGI, the assumptions on which these benefits rely need to be tested. Fortunately, Study 3 benefited from the opportunity to use ten high-fidelity CGI clips created by the same company that produced the CGI clips for the UK hazard perception test. By replicating the design of Study 2 (as closely as possible), we aimed to assess whether CGI clips were better at differentiating between driver groups when presented in a VR headset or on a single screen, and of course to see how the results compare with Study 2's video-based tests.

4.1.1 The current study

The ten CGI clips used by Crundall et al. (2021) were remade by the programmers (Jellylearn Ltd.) as 360-degree film clips, with additional assets inserted to create the wraparound world. They also provided a car interior with a driver's body (used in Study 2 as the car overlay). All clips were occluded at hazard onset, and participants were asked "What happens next?" Four text options were provided for participants to choose between.

As we had only ten CGI clips, the design of the study was modified to a between-groups comparison of their prediction scores. Novice and experienced drivers still completed both tests (counterbalanced across participants), but as they therefore saw all the hazards twice, it would have been unfair to compare their accuracy across these tests. For this reason, we compared drivers' performance on only the first test across the groups (comparing four groups' data from their first test: VR-experienced, VR-novice, single-screen-experienced, single-screen-novice).

The reason for subjecting participants to a second version of the test was to obtain CRIE responses. Although the participants were aware of the impending hazards on the second test (invalidating their behavioural responses), they were still able to compare CRIE levels across the two tests.

In other respects, the design remained the same as Study 2. We predicted that the single-screen clips should differentiate between a novice and an experienced driver group (following Crundall et al., 2021), though whether the VR version would be more or less successful than the single-screen test at separating these driver groups remained a non-directional hypothesis.

4.2 Method

4.2.1 Participants

The third study recruited 125 participants, split across experienced and novice drivers (64 experienced drivers and 61 novice drivers). Most novices were still learning to drive, though ten novices had passed their driving test within the 12 months prior to the experiment. Three experienced drivers and one novice driver who completed the 360-degree test first were removed from the experiment owing to their reported sickness levels reaching threshold.

Of those participants who undertook the VR test following the single-screen test, a further four experienced drivers were also removed from the experiment as a result of sickness, though we retained their behavioural response data from the first condition for analysis. This resulted in a total of eight participants who suffered from sickness in the study (seven experienced drivers and one novice), amounting to 6% of the sample.

A further three experienced drivers and two novice drivers were removed for various reasons (equipment failure, failure to give demographic information or misrepresentation of their driving status). The demographic details of the participants in each of the conditions and participants who suffered sickness are given in Table 4.1.

Table 4.1: Demographics of experienced and novice drivers in each condition who completed Study 3, showing participants who suffered from sickness in grey at the bottom of the table

Condition	N	Gender	Mean age (years)	Mean driving experience (years since passing driving test)
Single-screen – experienced*	31	16 females	43.4	21.1
Single-screen – novice	29	17 females	21.3	0.3
Virtual reality* – experienced	27	15 females	37.9	17.8
Virtual reality* – novice	29	14 females	18.8	0.1
*Single-screen – experienced***	*4*	*2 females*	*57.5*	*33.3*
Virtual reality – experienced	*3*	*3 females*	*54.9*	*30.3*
Virtual reality – novice	*1*	*1 female*	*25.2*	*0.5*

Source: Authors' own (Study 3)
Notes: * Not including participants who were removed due to sickness
** Including participants in the single-screen condition who suffered from sickness in the subsequent VR task

4.2.2 Design

A 2 × 2 between-groups design was employed to compare drivers' scores on the hazard prediction test across the factors of driving experience (experienced and novice drivers) and presentation mode (360-degree vs single-screen). With only ten hazards in the CGI test, we could not show different single-screen CGI hazards and 360-degree CGI hazards to the same participants without reducing the number of clips per condition to less than that used in Study 1. For this reason, presentation mode was treated as a between-subjects factor for the purpose of comparing behavioural and oculomotor data.

Regarding the CRIE questions, however, a 2 × 2 mixed design was retained: even though participants' behavioural data from their second test were not used for the analyses, participants were still able to provide CRIE ratings for both presentation modes (as these ratings are unlikely to be heavily influenced by having already seen the same clips in a different presentation mode).

4.2.3 Stimuli and apparatus

A single-screen CGI test had previously been designed, commissioned, and validated for a previous project for DfT (Crundall et al., 2021). The test consisted of a ten-minute drive through a CGI-rendered world, travelling on a variety of roads (arterial, suburban and rural). The video took the perspective of the driver, travelling through junctions, turning into side roads, and encountering ten pre-specified hazards (see Table 4.2 for a brief description of each hazard). The test was silent apart from a voice-over of an instructor providing guidance on where the film car would turn (e.g. "take the next left"). The ten-minute video was previously edited into ten clips, each stopping just as a hazard begins to onset. In the original study, Crundall et al. (2021) presented these hazard clips in a sequential order, creating the feel of a single drive interspersed with hazards.

The current 360-degree test was based on the clips used by Crundall et al., (2021). The routes and hazards remained the same, though the visual world was expanded to cover 360 degrees, allowing for the same visual flexibility that is contained in the 360-degree video test (Study 2). Additionally, the 360-degree CGI test included a car interior, side mirrors, a rear-view mirror, a digital speedometer embedded in the dashboard, and a driver's body in the footage, with hands holding the steering wheel. These new developments did not appear in the original clips used by Crundall et al. (2021). The same car interior was used in Study 2, though the arms of the 'driver' did not move in Study 2, whereas in the CGI clips the arms of the driver moved to turn the steering wheel in relation to the heading of the film car. The speedometer was also specific to Study 3 and did not appear in the video clips used in Study 2.

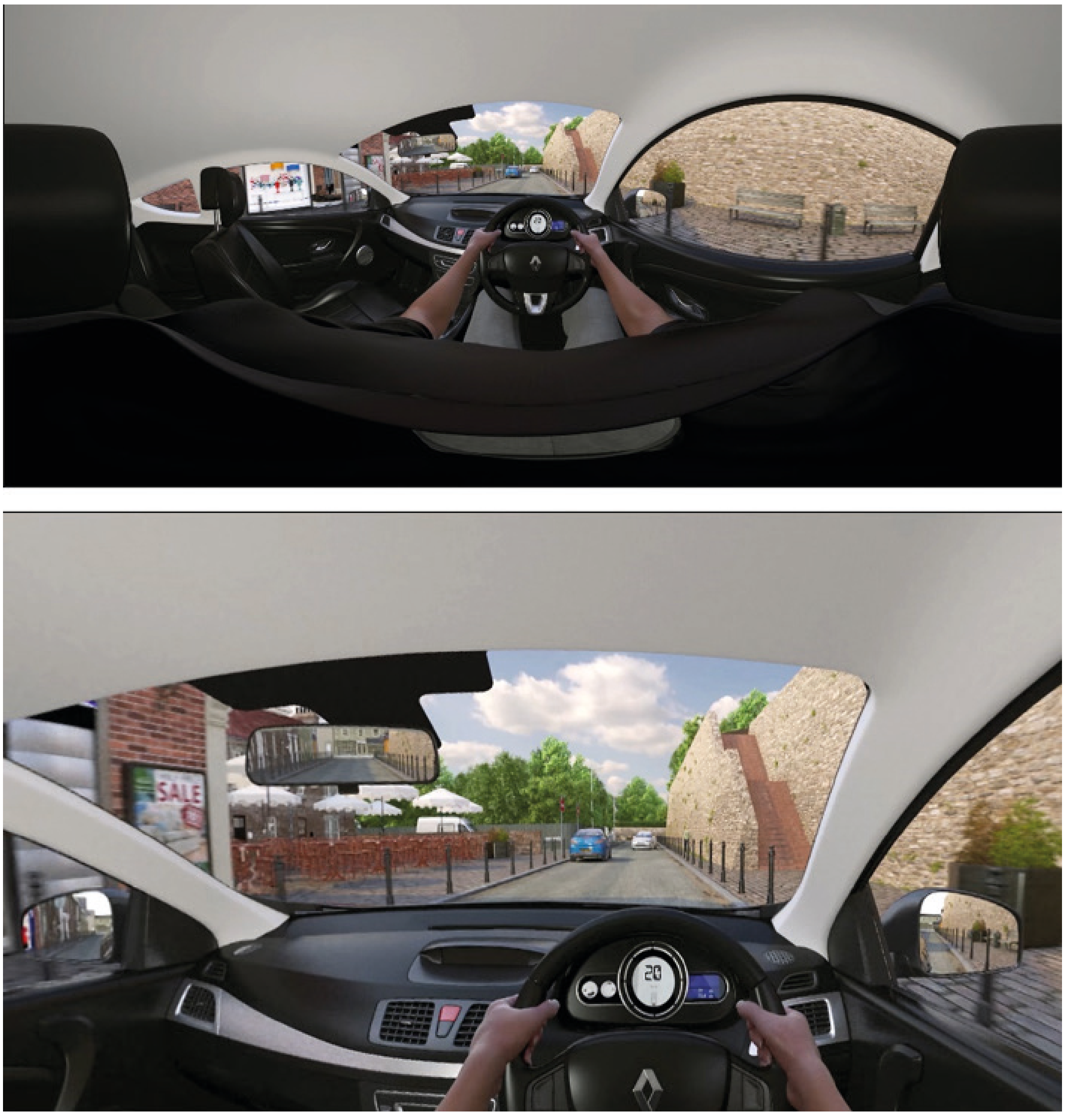
The single-screen variant of the CGI test was created in an identical process to that used to make the single-screen video-based test in Study 2 (see Figure 4.1 for a sample screenshot). As with Study 2 (and similarly to Crundall et al., 2021), the clips were occluded at the point of hazard onset and four options were presented for participants to choose between. The average overall clip length (this being also the time up to the occlusion point) across both clip sets was 59 seconds (range: 8 s to 159 s).

Other stimuli included questionnaires (demographics, SSQ and CRIE) and the practice clip. These were identical to those used in Study 2. The apparatus employed was also the same as that used in Study 2.

4.2.4 Procedure

The procedure was the same as that used in Study 2 in almost all regards. All instructions, the use of the long practice clip, and the points at which the SSQ was administered, were all identical to the previous study. The only difference was that the clips were presented sequentially to ensure the flow of the journey. This mirrored the approach used by Crundall et al. (2021).

Figure 4.1: Screenshot of the full 360-degree view of the CGI test created for Study 3 (top panel) and the cropped area used in the single-screen variant (bottom panel)



Source: Authors' own (Study 3)

Table 4.2: Description of the hazards in the test created for Study 3

Clip number	Description	Multiple-choice options (with correct answers underlined)	Clip length (sec)
1	An oncoming car turns across your path into a side road on your left. It is a one-way street with a van travelling in the opposite direction. The turning car is blocked and must reverse into your path.	1. The parked blue car on the left indicates and pulls off as you pass it. 2. The oncoming car turns into a side road, but must stop, blocking your way. 3. A white van pulls out of the side road on the left, forcing you to brake. 4. The oncoming car accelerates towards you, preventing you from overtaking the parked car ahead.	8
2	While you are travelling in the right lane of a two-lane carriageway, the car immediately ahead indicates and moves over into the left lane. Unfortunately, the driver fails to see a car in the left lane, hidden in their blind spot. The manoeuvring car narrowly misses the car in the left lane, but the latter driver pulls out immediately into the right lane to overtake. The overtaking manoeuvre of this second car is the hazard.	1. The red car in the left lane suddenly pulls into your lane. 2. The oncoming car turns sharply across your path in order to enter a driveway on your left. 3. The silver car ahead suddenly swerves back into your lane. 4. The silver car brakes harshly, forcing you to brake also.	70
3	You approach a crossroads intending to turn right. At the junction, an articulated lorry also intends to turn right, potentially obscuring oncoming traffic. As you make the turn, an oncoming motorcycle emerges from behind the lorry.	1. The LGV decides not to turn right and proceeds straight across the junction narrowly missing you. 2. A pedestrian steps into the road that you are trying to turn into. 3. An oncoming motorcycle prevents you from turning. 4. There is congestion on the road you are turning into, which forces you to stop.	39
4	You are driving along a narrow street with parked cars on either side. An oncoming car flashes its lights, as if to allow you through the bottleneck of parked vehicles. A second driver, visibly approaching from a side road, misinterprets this signal as an invitation to pull out. As you drive forward, the car suddenly emerges from the side road.	1. The passenger door of a car parked on the right suddenly opens. 2. A car emerges from a side street on the right, into your path. 3. A pedestrian steps into the road from between parked cars on the left. 4. The red car parked on the left indicates and pulls off in front of you.	60
5	As you are driving along a suburban route with infrequent parked vehicles, pedestrian movement can be noted through the windscreen of a parked car on the left. As you approach, a woman with a buggy almost steps out in front of you.	1. The white parked car on the left tries to pull off as you pass it. 2. A man carrying a large box steps out from behind a white van parked on the right. 3. An oncoming car turns across your path to enter a driveway on your left. 4. A woman pushing a buggy steps out from between parked cars on the left.	69

Clip number	Description	Multiple-choice options (with correct answers underlined)	Clip length (sec)
6	While you are driving along a suburban route, a cyclist can be seen on a cycle lane shared with the pavement on the left. The cyclist is travelling in the same direction as you, but you quickly pass them. The approach of a police car causes all vehicles to pull over briefly, which gives the cyclist time to catch up (though not visibly so). As you turn into a side road on the left, the cyclist crosses in front of you.	1. A pedestrian steps into the side road as you begin to turn. 2. As you attempt to turn, a car from right passes you heading for the same side road. 3. As you turn into the side road you find immediate congestion ahead that forces you to brake. 4. A cyclist crosses the side road as you begin to turn.	63
7	You are approaching a pedestrian crossing that has been on red for some time. As you slow down, a briefly visible pedestrian, mostly occluded by a parked car, decides to cross the road. The lights change and you are about to accelerate when the pedestrian emerges.	1. A pedestrian runs into the road from the left from behind a parked car. 2. The lights at the pedestrian crossing turn red, forcing you to stop. 3. The blue car parked on the left suddenly indicates and tries to pull off in front of you. 4. The car ahead suddenly brakes, forcing you to brake also.	24
8	When you are trying to overtake a stationary bus, a car can be briefly seen approaching from a side road on the left, ahead of the bus. As you pass the bus, the car pulls out of the side road.	1. The bus indicates and starts to pull off as you attempt to pass it. 2. A pedestrian emerges from in front of the bus on the left. 3. A car emerges from a side road on your left. 4. The oncoming car accelerates towards you, preventing you from overtaking the bus.	66
9	A zebra crossing precedes a mini-roundabout ahead. A pedestrian from the left crosses in good, time, but a pedestrian on the right crosses in front of you. His intention to cross is signalled by a change in trajectory and a glance at your car, but an oncoming vehicle then obscures him. After this vehicle passes, the pedestrian appears on the crossing in front of you.	1. The oncoming car strays into your lane. 2. A car enters the mini-roundabout ahead from the right. 3. A pedestrian crosses the road from the right. 4. A car enters the mini-roundabout from the left.	35
10	There is a standing line of traffic in the oncoming lane. You intend to turn into a side road on the right. A car approaches slowly from this side road but does not pose a threat. Instead, an oncoming motorcycle decides to overtake the standing traffic just as you try to make the turn.	1. An oncoming motorcycle prevents you from turning. 2. One of the cars waiting in the oncoming lane closes the gap into the side road, preventing you from turning. 3. A red car emerges from the side road on the right and pulls out in front of you. 4. A pedestrian steps out from between waiting cars on the right.	159

Source: Authors' own (Study 3)

Note: Where descriptions include "You intend to…", this refers to a voice-over that tells the driver where the car will turn next.

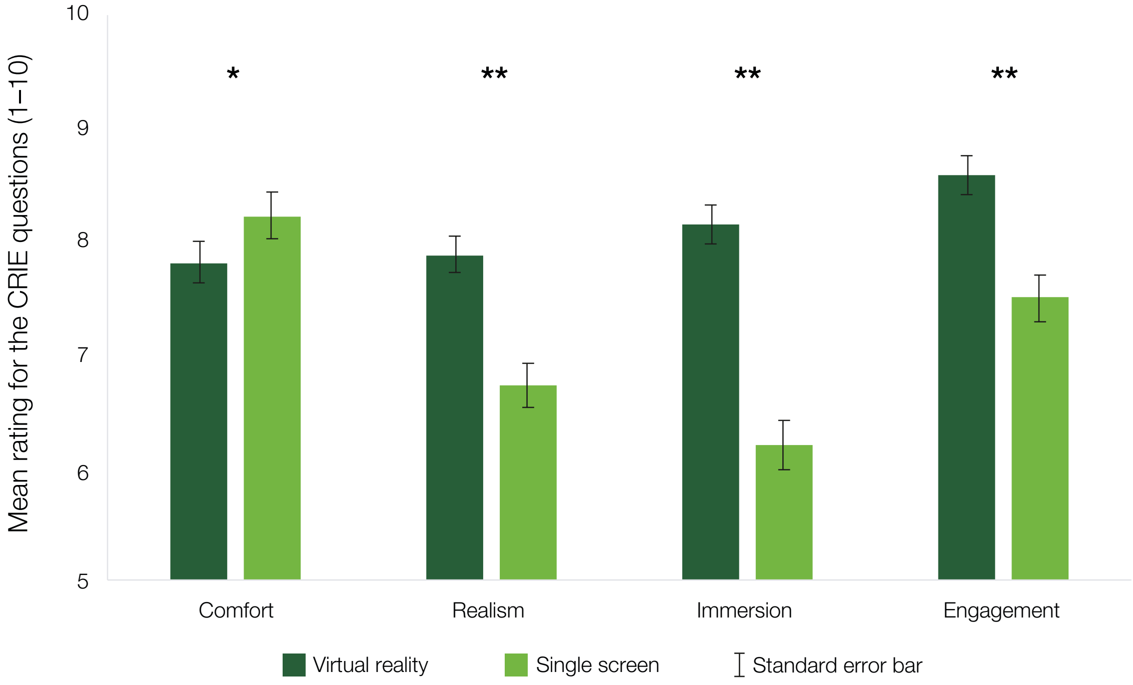
4.3 Results

4.3.1 Comfort, realism, immersion, and engagement questions

As in Studies 1 and 2, participants gave ratings for each of the questions regarding CRIE for both the VR and single-screen tests. Although behavioural performance measures for each participant were taken only from the first test they completed, the CRIE ratings were recorded for both tests, as these were considered unlikely to be strongly affected by whether they had already seen the hazards. The seven experienced drivers and the one novice driver who were removed owing to sickness were not included in this analysis. A further four participants who did not complete the second session were also not included (though their behavioural data from their first session was included in further analyses).

Participants' ratings were entered into a series of 2 × 2 between-groups ANOVAs across the factors of driver group and presentation mode. Main effects of presentation mode were found for all four ratings, with drivers reporting the 360-degree test to be less comfortable (7.8 vs 8.2), but having greater realism (7.9 vs 6.7) and immersion (8.1 vs 6.2), and evoking greater engagement (8.6 vs 7.5) than the single-screen test (comfort: $F(1, 106) = 4.7$, $MSE = 1.9$, $p = .03$, $\eta^2 = .04$; realism: $F(1, 106) = 36.9$ $MSE = 1.9$, $p < .001$, $\eta^2 = .26$; immersion: $F(1, 106) = 85.6$, $MSE = 2.3$, $p < .001$, $\eta^2 = .45$; engagement: $F(1, 106) = 28.3$, $MSE = 2.2$, $p < .001$, $\eta^2 = .21$). The means are displayed in Figure 4.2.

Figure 4.2: Average ratings given by experienced drivers for each of the four items on the Study 3 CRIE questionnaire for the virtual reality and single-screen tests



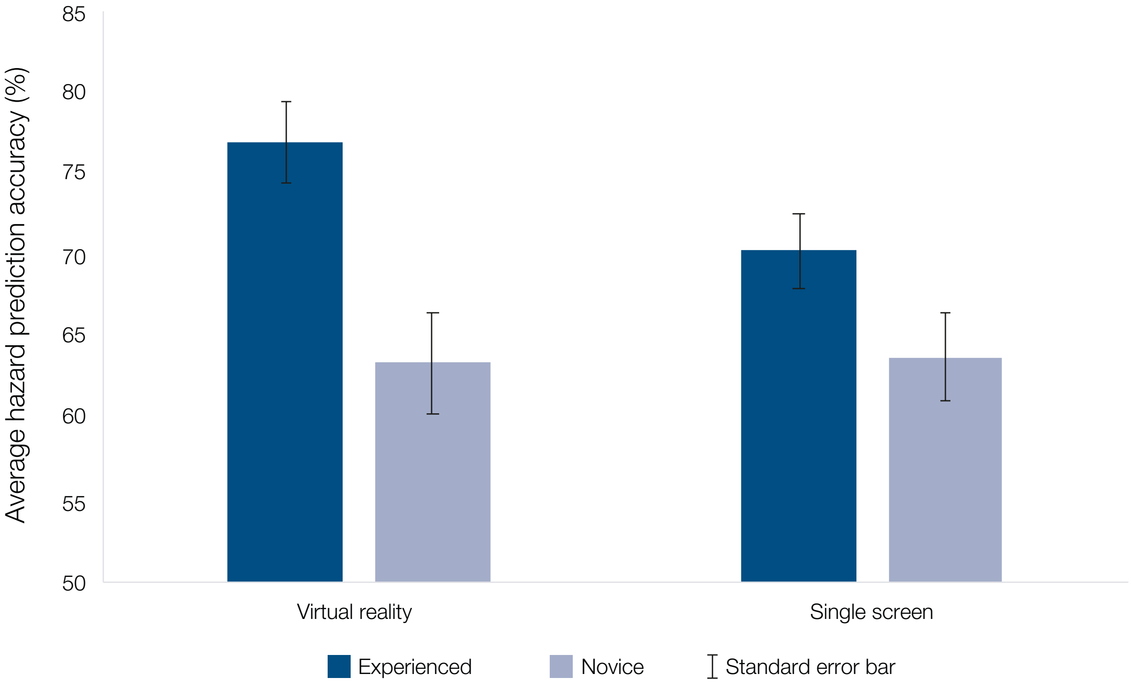
Source: Authors' own (Study 3)
Note: *p < .05, **p < .001

4.3.2 Hazard prediction performance

All participants saw the same ten clips in both the VR headset and on the single screen. Only the data from the first presentation mode which they encountered was included in this analysis. The percentage of hazards that they correctly predicted was entered into a 2 × 2 between-groups ANOVA across driver group and presentation mode. This analysis revealed a main effect of driver group ($F(1, 112) = 14.0$, $MSE = 208.9$, $p < .001$, $\eta^2 = .11$), with experienced drivers predicting significantly more hazards than novice drivers (73.4% vs 63.6%; see in Figure 4.3). Presentation mode did not produce a significant effect. While Figure 4.3 suggests the presence of an interaction, this did not reach the threshold of statistical significance, $F(1, 112) = 1.7$, $MSE = 208.9$, $p = .19$, $\eta^2 = .02$.

In a similar pattern of results to those of Study 2, the evident improvement in differentiating the driver groups in the VR condition is supported by the pre-planned comparisons which show that experienced drivers outperformed novice drivers (77.0% versus 63.4%, $p = .001$) while viewing the VR clips. In the single-screen condition, although experienced drivers scored more highly than novices (70% versus 64%), this difference did not reach the threshold for significance ($p = .07$). Specifically, the results suggest that the 360-degree test allows the experienced drivers to better demonstrate their hazard skills. However, the failure of this effect to result in a significant interaction weakens the evidence of the pre-planned comparisons.

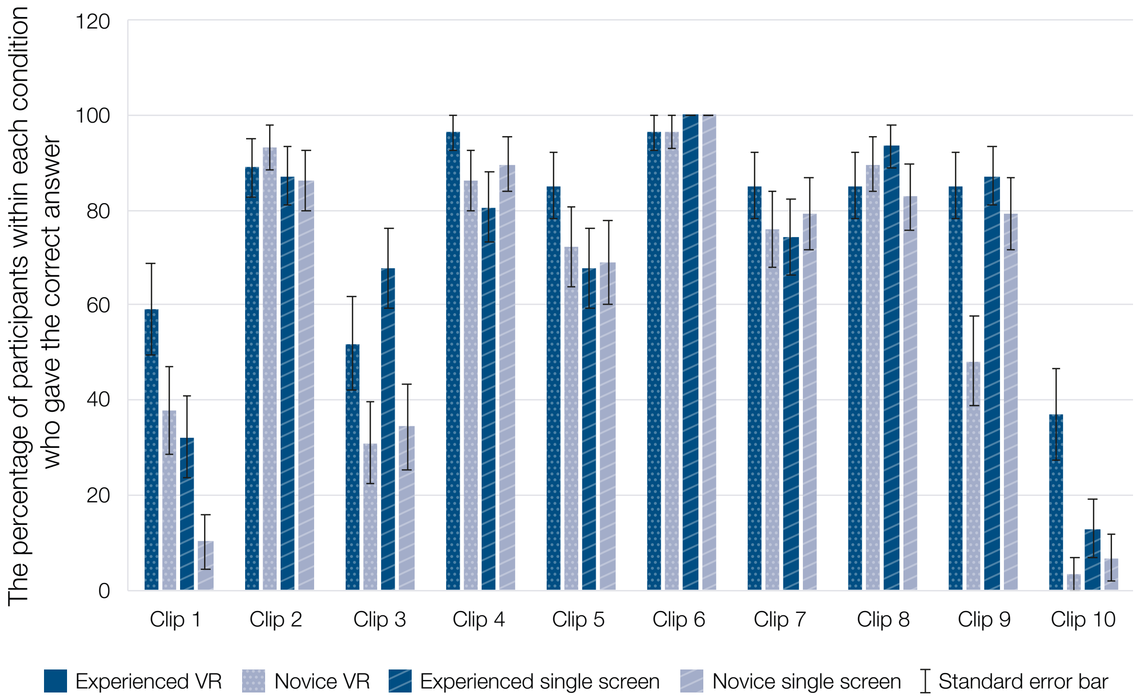
Figure 4.3: Hazard prediction performance for the virtual reality and single-screen tests of Study 3



Source: Authors' own (Study 3)

To assess the contribution of individual clips to this effect, the accuracy for each clip, for each experience group and for each presentation type (VR or single-screen) was also charted (Figure 4.4). As noted in Study 2, the pattern of correct responses changes across the clips (as was also noted in the original study by Crundall et al., 2021). Several clips show a clear advantage of the VR presentation mode, though possibly for different reasons. For instance, the pattern of responses for clip 9, suggests that experienced drivers did not perform better when in the VR headset compared to the single-screen condition. Instead the evidence in favour of the VR condition results from a deterioration in performance from the novices in VR compared to the single-screen condition. This suggests that the 360-degree environment allowed novices to look at areas of the scene other than where the hazard was about to appear from. Looking at clip 10, however, the opposite pattern is observed in experienced drivers. Novice drivers remain resolutely poor on this clip, whereas the experienced drivers show a threefold improvement in the VR condition compared to the single-screen condition.

Figure 4.4: Hazard prediction accuracy scores across all clips for the virtual reality and single-screen tests of Study 3



Source: Authors' own (Study 3)

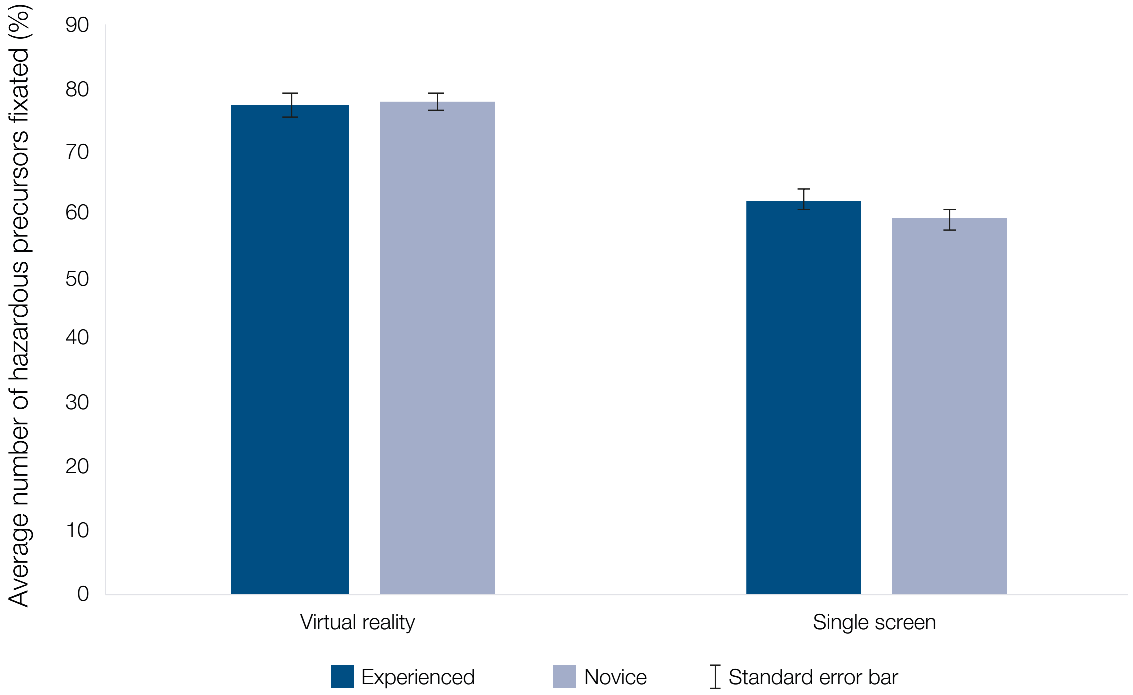
4.3.3 Eye-movement measures

As in Study 2, several eye-movement measures were calculated to reflect whether participants looked at each hazard, and if so, how quickly they looked at them, and for how long. In all analyses of eye data, 11 participants were removed owing to poor calibration in the single-screen test.

4.3.3.1 Did they spot the hazards?

The percentage of hazardous precursors that drivers (both novice and experienced) fixated was calculated (out of ten). In addition to those participants removed, a further 11 participants in the single-screen condition showed poor calibration for an average of 1.5 hazard clips. For these participants, the percentage of hazard precursors they fixated was calculated out of the total number of hazard windows for which they had sufficient eye data. This data was subjected to a 2 × 2 between-groups ANOVA across driver group and presentation mode. This analysis revealed a significant main effect of presentation mode ($F(1, 101) = 27.8$, $MSE = 259.8$, $p < .001$, $\eta^2 = .22$), with those participants who saw the test in the VR headsets fixating significantly more of the hazardous precursors than those who saw them on a single screen (77.3% vs 60.6%; see figure 4.5).

Figure 4.5: Average percentage of hazards that participants looked at in the virtual reality and single-screen tests of Study 3

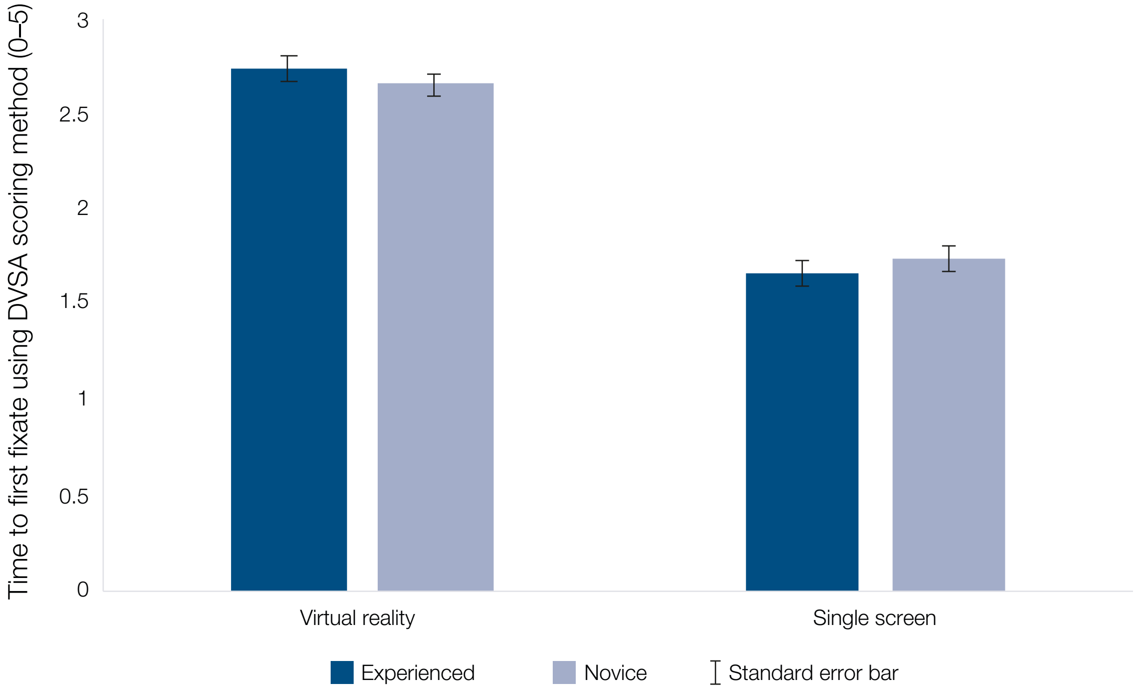


Source: Authors' own (Study 3)

4.3.3 How soon do participants fixate the hazard precursors?

Participants' time to first fixate the hazardous precursors was calculated as the time at which participants first looked at the hazard minus the hazard onset time. The DVSA scoring method was applied to time-to-first-fixate measure (see subsection 3.3.3.2). The data was compared across driver group and presentation mode in a 2 × 2 between-groups ANOVA. This revealed a main effect of presentation type ($F(1, 101) = 58.3$, $MSE = 0.5$, $p < .001$, $\eta^2 = .37$). Participants who saw the test in the VR headset fixated hazards faster than those who saw them on a single screen (2.7 points vs 1.7 points; see Figure 4.6). There was no difference between driver groups, and the interaction was not significant.

Figure 4.6: Average number of points scored, reflecting how fast participants fixated the hazardous precursor in each test type in Study 3



Source: Authors' own (Study 3)

4.3.3.3 Amount of attention devoted to the hazard precursors

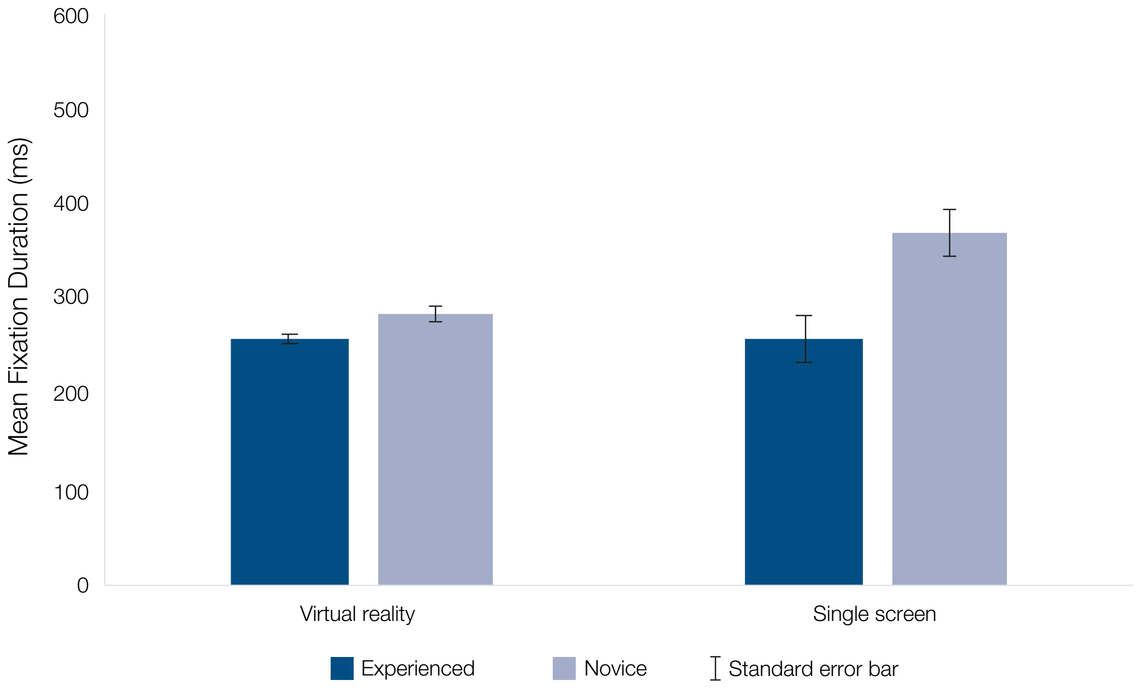
In line with Study 2, we analysed the measures of mean fixation duration, first-fixation duration and dwell time chosen to reflect attention given to the hazard precursors (often used to give an indication of the level of cognitive difficulty that drivers encounter when processing hazards).

For mean fixation duration (see Figure 4.7), the average amount of time participants spent looking at the hazardous precursors was calculated for both test types. This was compared across driver group and presentation mode by means of a 2 × 2 between-groups ANOVA. This revealed a main effect of experience ($F(1, 101) = 5.6$, $MSE = 31,712.5$, $p = .02$,

$\eta^2 = .05$), with experienced drivers having significantly shorter mean fixation durations on the hazardous precursors than the novice drivers (289 ms vs 367 ms). This reflects the experienced drivers' greater ability to process driving-related information and has been noted in many previous studies (e.g. Chapman & Underwood, 1998), though it should be noted that the effect size is small.

There was also marginal evidence for an effect of presentation mode ($F(1, 101) = 3.7$, $MSE = 31{,}712.5$, $p = .057$, $\eta^2 = .04$), suggesting that participants in the single-screen condition had longer fixation durations than participants in the VR condition (365 ms vs 297 ms). Despite an ostensible trend towards an interaction, it did not reach standard levels of significance ($F(1, 101) = 2.9$, $MSE = 31{,}712.5$, $p = .09$, $\eta^2 = .03$).

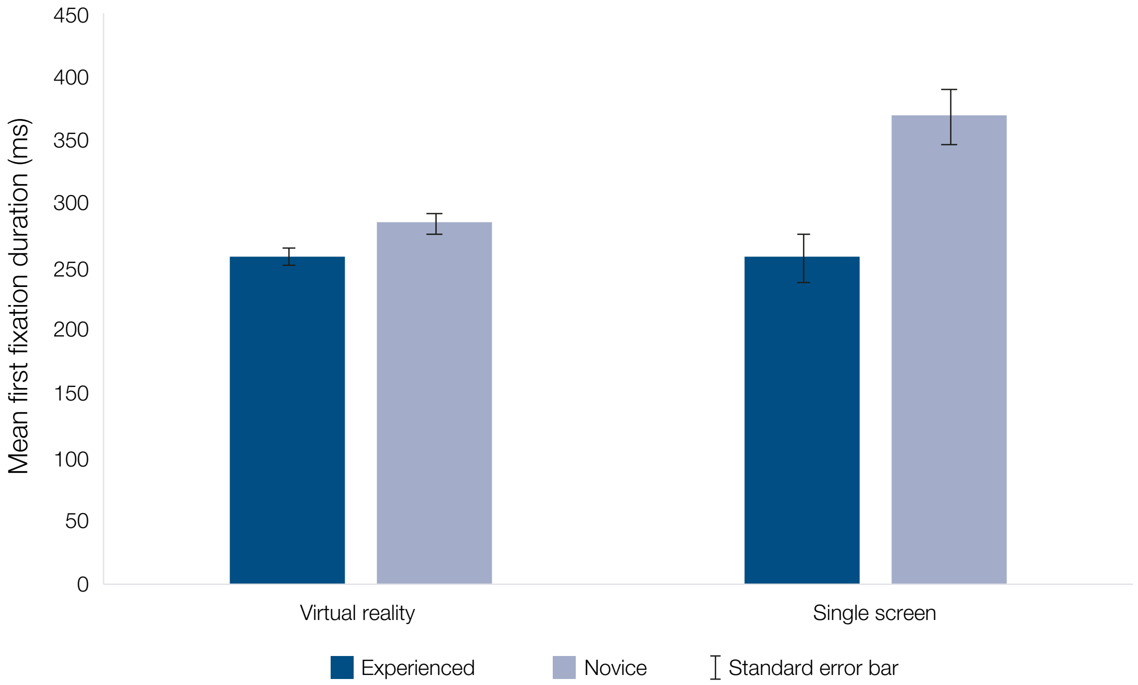
Figure 4.7: Average fixation duration for each group on the hazardous precursors for each test type in Study 3



Source: Authors' own (Study 3)

The average duration of the first fixation on the hazardous precursors was also calculated for both test types and subjected to a similar analysis. A main effect of experience ($F(1, 101) = 5.2$, $MSE = 23{,}731.8$, $p = .03$, $\eta^2 = .05$) revealed experienced drivers to have significantly shorter first-fixation durations on the hazardous precursors than the novice drivers (258 ms vs 323 ms; see Figure 4.8). This again suggests that experienced drivers are better able to extract visual information even within the first fixation on a hazardous precursor. The presentation mode did not affect first-fixation durations, and the interaction was not significant, ($F(1, 101) = 2.0$, $MSE = 23{,}731.8$, $p = .16$, $\eta^2 = .02$), despite a visual trend similar to that seen in Figure 4.7.

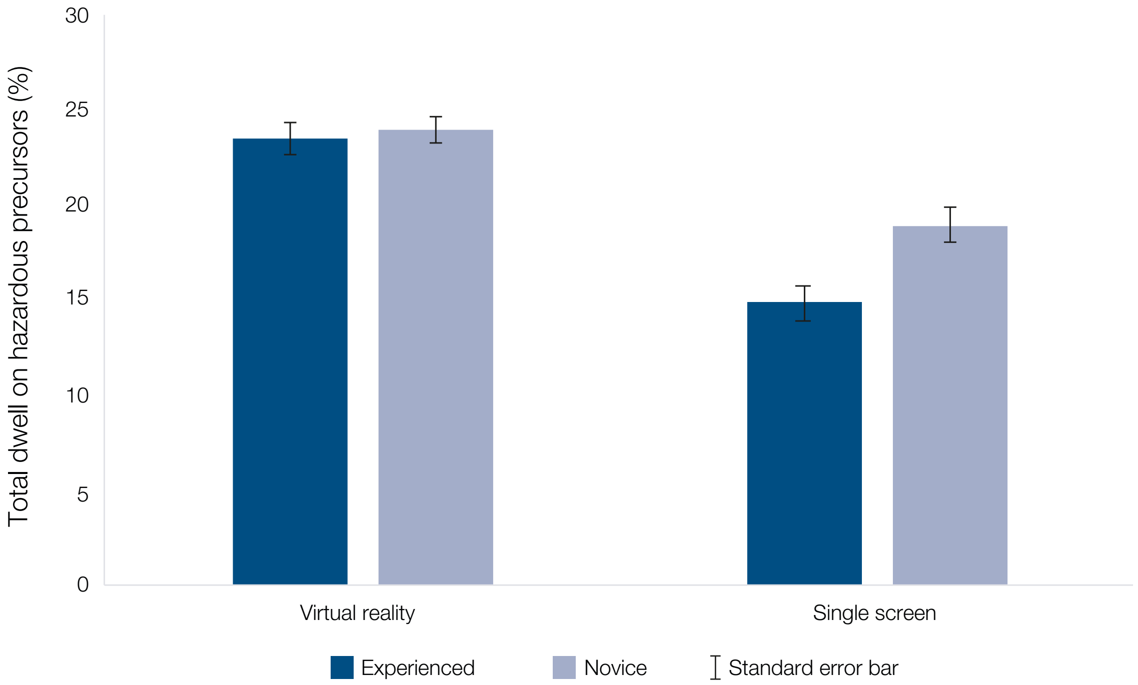
Figure 4.8: Average first-fixation duration for each group on the hazardous precursors in Study 3



Source: Authors' own (Study 3)

Finally, the dwell-time measure was calculated as the percentage of time the participants spent looking at the hazardous precursor as a function of the amount of time it was visible on-screen. This was compared across both driver groups and presentation modes using a 2 × 2 ANOVA. There was a significant effect of presentation mode ($F(1, 101) = 16.5$, $MSE < .01$, $p < .001$, $\eta^2 = .14$). As with Study 2, participants in the VR condition spent longer looking at the hazardous precursors than those participants who viewed the test on the single screen (23.7% vs 16.9%; see Figure 4.9). There was no effect of driver group and the interaction was not significant (both values of $p > .05$).

Figure 4.9: Average dwell time (%) on the hazard precursors in Study 3 across the different participant groups



Source: Authors' own (Study 3)

4.3.3.4 Mirror and dashboard usage

One ancillary question of interest was whether group differences could be identified in their visual inspection of the internal elements of the car – namely the three mirrors and the dashboard (which includes the speedometer). Inexperienced drivers have been found to inspect mirrors less frequently than more-experienced drivers (Lee, Olsen & Simons-Morton, 2006; Underwood, Crundall & Chapman, 2002), yet they still have greater eyes-off-road durations as a result of in-vehicle distraction, such as engaging with dashboard-mounted devices (e.g. Simons-Morton et al., 2014). We wanted to ascertain whether the experiential effects noted in other studies are replicated with the current stimuli.

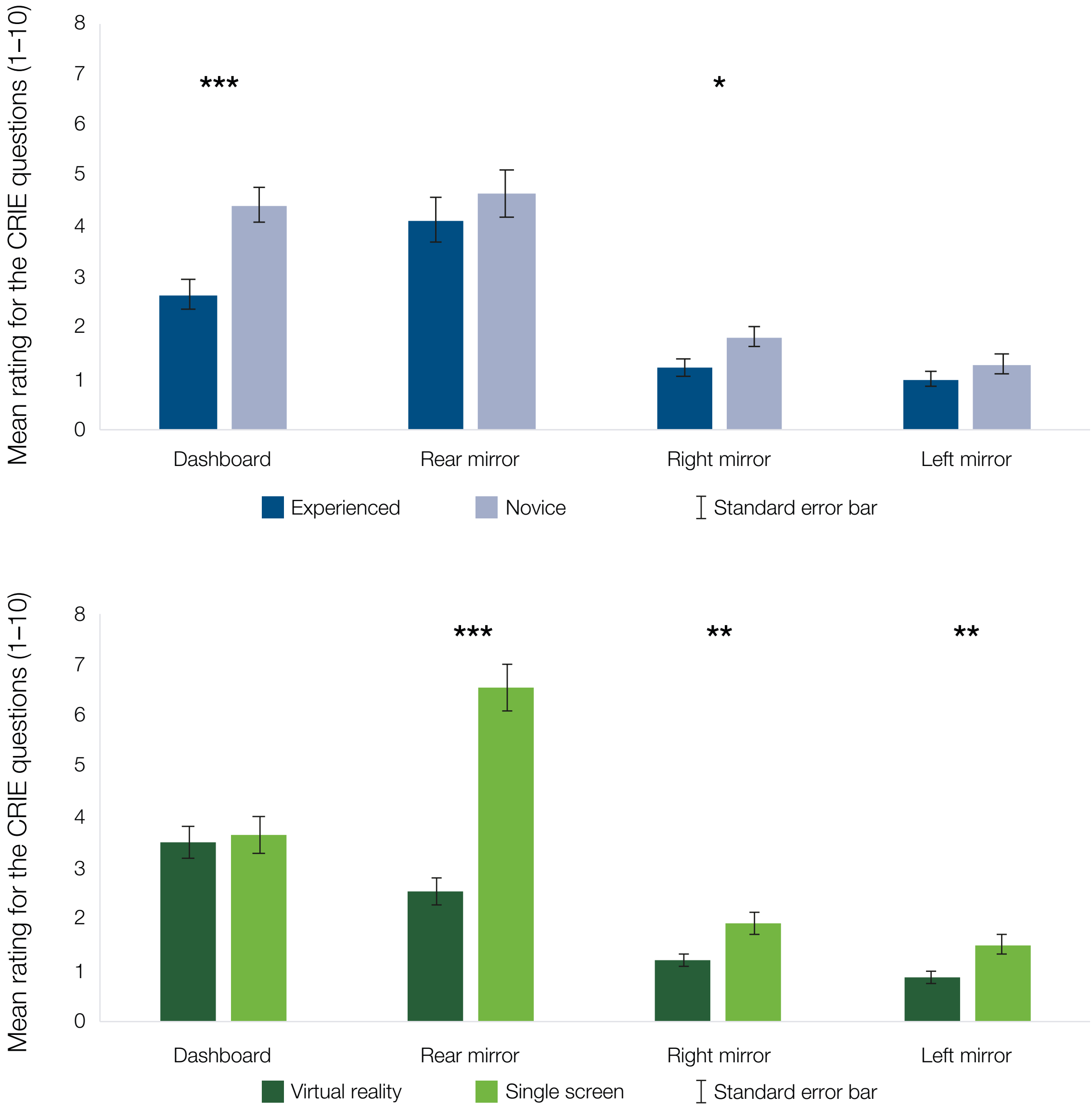
The dwell time of participants on the four AOIs (the three mirrors and dashboard) was calculated and compared across driver groups and presentation mode by means of a 2 × 2 × 4 mixed ANOVA.

All main effects were significant. Regarding the effect of AOIs ($F(3, 303) = 85.6$, $MSE < .01$, $p < .001$, $\eta^2 = .46$), the dashboard and rear-view mirror each received approximately 4% of the available inspection time (3.9% and 4.1%, respectively) while the right- and left-side mirrors received significantly less attention (1.5% and 1.1%, respectively). The effect of driver group ($F(1, 101) = 11.9$, $MSE < .01$, $p = .001$, $\eta^2 = .11$) found novice drivers ($M = 3\%$) spent more time looking at the four AOIs than experienced drivers ($M = 2\%$). Finally, the main effect of presentation mode ($F(1, 101) = 35.9$, $MSE < .01$, $p < .001$, $\eta^2 = .26$) suggested that the single-screen test evoked more gaze on the internal AOIs than the VR test (3.3% vs 2.0%).

These main effects were, however, subsumed by two interactions. The first significant interaction was between AOI and driver group (*F*(3, 303) = 3.2, *MSE* < .01, *p* = .02, η^2 = .03). Post hoc corrected t-tests revealed that regardless of presentation type, novice drivers spent more time looking at the dashboard than the experienced drivers (4% vs 3%, *p* < .001). The novices also spent more time looking at the right-side mirror than the experienced drivers (2% vs 1%, *p* = .02; see Figure 4.10, top panel).

The second significant interaction was across presentation mode and AOI (*F*(3, 303) = 25.55, *MSE* < .01, *p* < .001, η^2 = .20; Figure 4.10, bottom panel). Post hoc corrected t-tests revealed that the single-screen test evoked longer dwell on all mirrors than did the VR test (rear-view mirror, 7% vs 3%, *p* < .001; left-side mirror, 2% vs 1%, *p* < .005; right-side mirror, 2% vs 1%, *p* < .005).

Figure 4.10: Average dwell time (%) within different AOIs in Study 3 across driver group (top panel) and presentation mode (bottom panel)



Source: Authors' own (Study 3)
Note: * p < 0.05; ** p < 0.005, *** p < 0.001

Several conclusions can be drawn from this analysis. First, the dwell time on the mirrors reflects their differing levels of saliency. The rear-view mirror is closer to the visual heading of the vehicle and is also the largest of the three mirrors. It also contains more information than the side mirrors and is therefore a more useful source of information except in very specific circumstances in which the driver is looking for objects that are positioned to the rear near-side and rear off-side.

The increased size of the scene in the VR headset means that the participant must move their head to comfortably fixate the mirrors (especially the left-side mirror). Given the extra effort required to look in the mirrors in the VR condition (which better emulates the level of real-world effort), it is understandable that this presentation mode results in lower dwell times. The fact that the rear mirror is most impacted in the VR condition may be additionally influenced by the ease with which participants could fixate the rear-view mirror in the single screen condition.

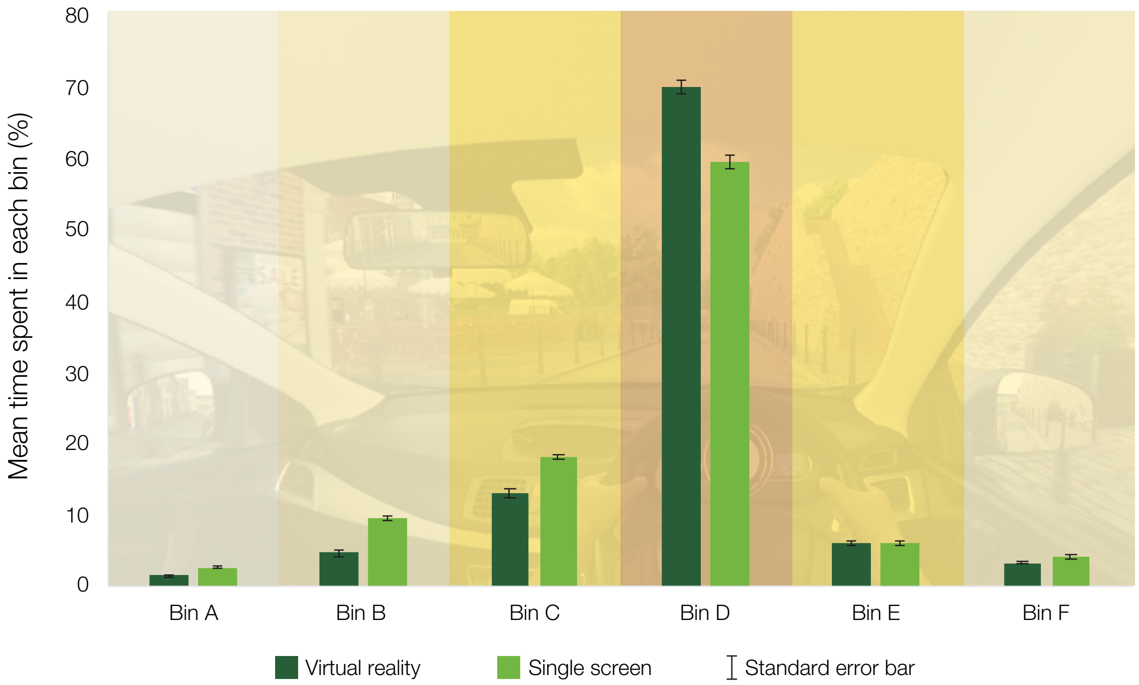
Between 3% and 4% of time was spent looking at the dashboard (e.g. the speedometer). The longer dashboard dwell time for the novices fits with longer eyes-off-road times noted previously (e.g. Simons-Morton et al., 2014), though novices' greater dwell on the right-side mirror is an interesting finding that is not typical of this field. It is possible that these drivers are following explicit guidance given to them by their driving instructors, which might be rejected once they have passed.

4.3.3.5 Spread of search analyses

A second ancillary analysis was undertaken on the spread of eye movements across the visual scene, comparing the driver groups across the two presentation modes. Six bins were created to capture dwell on different areas of the single-screen test (covering 95% of the available visual information from the left-side mirror to the right-side mirror). The same bins were applied to the 360-degree clips. While this potentially excluded more eccentric eye movements in the VR condition (as these edge areas would not feature in the single-screen view), this was considered a fairer comparison between the two tests. Dwell for each bin was calculated as a percentage of the total time spent in all the bin areas. Any time that the eyes spent outside these six bins was classified as 'other'. The other category amounted to 1% of all dwell in the VR test and <1% in the single-screen test. Given these small percentages, they were not included in the bin analysis.

Analysis of dwell time in these bins (labelled A to F in Figure 4.11) revealed a significant interaction between presentation mode and bin (*F*(5, 505) = 49.5, *MSE* < .01, *p* < .001 η^2 = .33). Post hoc corrected t-tests revealed participants in the VR headset spent less time looking at bins A (2%), B (5%) and C (13%) than the participants in the single-screen condition (3%, 10%, and 18%, respectively). However, participants in the VR headsets spent more time looking at Bin D than participants in the single-screen condition (70% vs 60%; see Figure 4.11). No other differences were observed.

Figure 4.11: Percentage of time all participants spent looking at each bin in Study 3, regardless of driver experience



Source: Authors' own (Study 3)

The binned data suggests that, despite the additional opportunities to scan more widely in the 360-degree driving clips, participants in the VR condition actually concentrated their visual search more in Bin D than the participants in the single-screen condition. This suggests an apparently paradoxical notion that placing participants in a VR headset reduces the spread of visual search.

In truth, however, drivers still increase their absolute visual search in the VR condition (in terms of the angles through which their eyes travel). To look from one side mirror to the other in VR requires an angular change of over 100 degrees, whereas in the single-screen test this is approximately 40 degrees. Significantly, however, the increased viewing opportunities available in the VR condition do not result in drivers seeking out more eccentric information than they could get in the single-screen test.

4.4 Discussion

The primary aim of Study 3 was to identify whether a 360-degree hazard test, created with CGI, would be viewed more favourably by participants than a single-screen version, and whether it would be better able to differentiate between novice and experienced drivers. The data revealed that all drivers thought the test presented in the VR headset was more immersive, more realistic and more engaging than its single-screen counterpart. They did, however, rate the VR test as less comfortable than the single-screen test, though the effect size noted with this difference was much smaller that the effect sizes for realism, immersion and engagement (with partial eta squared values of 0.04, 0.26, 0.45, 0.21, respectively). This data mirrors that found in Study 2 with the naturalistic video-based test, with the exception of the worse comfort ratings for the VR CGI test.

In line with the results from Study 2, the results from Study 3 also showed that experienced drivers were able to successfully predict more upcoming hazards than the novice drivers. This finding is in line with the literature that argues that experienced drivers have better hazard prediction skills than novice drivers (Jackson et al., 2009; Castro et al., 2014; Crundall, 2016; Lim et al., 2014; Ventsislavova et al., 2016, 2019; Crundall & Kroll, 2019).

Figure 4.3 hinted at an interaction between presentation mode and driver group, with the suggestion that the greater difference between novice and experienced drivers was found in the VR condition. The interaction did not reach the threshold for statistical significance, though the pre-planned comparisons between the two groups for each test found that only the 360-degree clips produced a significant difference between experienced and novice drivers.

Also consistent with Study 2, the eye data analyses revealed no differences between experienced and novice drivers in terms of how fast they detected the hazards, as well as no difference in the number of precursors they looked at. As in Study 2, this suggests that, although the novice drivers were looking at the same number of hazardous precursors, they were not extracting sufficient information from the precursor to allow them to correctly predict the hazard (Crundall et al., 2012a).

Study 3 did, however, find experienced drivers to have shorter fixation durations on the precursors. This finding is consistent with previous literature that links shorter fixations to driving experience. It has been argued that experienced drivers' exposure to similar situations over years of driving allows current information to be processed rapidly, and irrelevant information to be disregarded faster (e.g. Chapman & Underwood, 1998; Crundall & Underwood, 1998). But the question arises as to why this effect was found in the CGI test, but not with the video-based test used in Study 2.

There are several possible reasons. For instance, Study 3 used hazards designed by experienced drivers (the team of traffic psychologists). These hazards may have contained implicit biases that made them more easily spotted by other drivers with similar levels of experience. In contrast, Study 2 contained natural hazards, free of any bias in their design. These natural hazards may have been free from any artificial fixation duration benefit later seen in Study 3.

Alternatively, the increased complexity and visual clutter that is available in the video-based hazards may create uncertainty in the outcome of the unfolding events.

In such situations, an experienced driver may process a potential hazard, but then remain fixated in order to confirm their suspicions. Thus, experienced and novice drivers could produce similar length fixations on natural hazards but for different reasons, belying the experienced drivers' faster processing of such precursors. Conversely, the reduced complexity of the CGI clips (with fewer distracters and less visual clutter), and the archetypal nature of the designed hazards, may better allow the application of 'mental templates' or exemplars of similar situations that are stored in long-term memory (Crundall, 2016; Pammer & Blink, 2013). If such templates can be more easily accessed and matched to the current situation, this may explain why the experienced drivers have shorter fixations on these clips (Vlakveld et al., 2011).

A further inconsistency with Study 2 is that, although the participants in Study 3 did look at more precursors in the VR than in the single-screen condition, they did not find the VR hazards easier to predict. This again is possibly linked to the relative simplicity of the CGI clips compared to the video-based clips. As the video clips are visually complex and likely to contain many subtle cues to the upcoming hazards, the increase in visual size in the VR condition will improve drivers' ability to identify and isolate the important cues. In the CGI clips, however, the hazardous precursors may be quite visible regardless of their size, owing to the less visually cluttered scenarios and backgrounds.

Two additional analyses were also conducted on the Study 3 data. The first looked at participants' gaze directed towards mirrors and the dashboard. The results showed that regardless of the test type, whilst all drivers looked at the dashboard and rear-view mirror more than the other mirrors, novice drivers spent more time looking at the dashboard and the right-side mirror than the experienced drivers. This suggests that the novices were spending more time looking away from the road than the experienced drivers, which may in part explain why they were not as accurate as experienced drivers in predicting the upcoming hazards. The analysis also revealed that regardless of driver experience, all three mirrors received more attention in the single-screen test than in the VR test.

This is corroborated by the comparison of drivers' spread of visual search. From an analysis of binned dwell times, we found that drivers in the VR condition restricted their visual search in relative terms. Thus, though our participants report greater immersion, realism and engagement resulting from the use of 360-degree clips, they make less use of the available information than in the single-screen condition. It is possible that the additional effort of moving the eyes in VR (because of the greater screen size) makes drivers feel as if they are engaged in a relatively active search of the scene, when in truth they are adopting a more concentrated visual search than in the single-screen test.

This may be explained by the available visual field. Specifically, in the VR test they have more of the scene to examine (i.e. they can turn their heads to look for hazards), whereas in the single-screen test there is much less of the visual field available and participants may therefore rely more on the mirrors to look for hazards.

In conclusion, as with Study 2, both tests showed an ability to discriminate between the driver groups, with a weak effect to suggest that VR might be most effective at this. However, the VR test is clearly a favourite amongst participants on several of the other subjective measures, and at any rate does not reduce the effectiveness of discriminating between the driver groups.

5. Study 4: Investigating a Training Benefit in a 360-Degree Environment

5.1 Introduction

The findings from Studies 1–3 have demonstrated a strong participant preference for the VR test, with realism, immersion and engagement being rated as higher than single-screen equivalents. While there are some understandable differences in the comfort associated with the two presentation modes, drivers appear willing to endure minor discomfort for the additional CRIE benefits. Severe discomfort appears low, with only 3% of participants removed from Study 2, and 6% removed in Study 3 on account of it. The behavioural data is also promising, demonstrating that the VR tests can differentiate driver groups on the basis of driving experience. There is even weak evidence to suggest that the VR tests can be more effective than 2D ones at identifying safer/more-experienced drivers. The eye-movement data provides mixed results, with one of the most interesting findings being the suggestion that the opportunity to scan beyond the confines of a single screen does not necessarily mean that one will seek out more eccentric information. Experiential eye-movement differences were more noticeable in the CGI test, and the possibility that these effects were due to the less-complex and less-cluttered CGI environment has been discussed. Regardless of the vagaries of the oculomotor evidence, the overall findings from the studies are very promising for the future of VR hazard assessment in terms of public acceptance, sickness levels and behavioural measures.

Assessment of drivers' hazard prediction skills is, however, only the first step in reducing collisions that are caused by a lack of hazard awareness. While an assessment test can be used to identify those drivers with a greater crash risk than others, we should then seek to improve the hazard awareness skills of those poorly performing drivers. Theoretically, a 360-degree environment should provide a better setting in which to train drivers in hazard prediction, as it more closely resembles the real world in several aspects. For instance, the need to move the head to fixate the left-side mirror can create motor memory that encourages mirror inspections in the real world. Viewing the left-side mirror in a single-screen representation, with the unrealistically small head movement required, could arguably be counterproductive, as drivers may later feel uncomfortable rotating the head and moving the eyes over a greater distance to reach the mirror.

5.1.1 Training hazard awareness

Evidence for the positive benefits of training on driver safety is relatively rare in the literature; however, hazard perception skill does appear to be one of the more promising targets for such interventions (Shinar, 2007). Several studies have demonstrated improvements in this skill through a variety of different training methods. For example, several studies have shown that listening to an expert commentary can have a positive impact on hazard perception (Castro et al., 2016; McKenna et al., 2006; Wallis & Horswill, 2007; Crundall et al., 2010; Horswill et al., 2010; Horswill et al., 2013; Isler et al., 2009; Poulsen et al., 2010; Wetton et al., 2013). Furthermore, various iterations of the RAPT (Risk Awareness and Perception Training) programme have used explicit instruction and exercises (e.g. clicking on AOIs to identify areas of the scene that require particular attention) to teach drivers where to look to spot hazards (Fisher et al., 2006, 2007, Pradhan et al., 2005, 2009). A study in California suggests that the latest version of RAPT has significantly reduced collisions of trained drivers, compared to a control group, in a large randomised control study (Thomas et al., 2016). A similar approach was adopted by Chapman et al. (2002), who used animated ellipses of changing size and colour, overlaid on hazard perception clips, to demonstrate where safe drivers should look. They found that drivers consequently had improved visual scanning on real roads, with some effects lasting several months after training. Other interventions have provided alternative visual perspectives on hazard scenarios, using plan views to encourage consideration of areas of the scene that might be occluded from a driver's typical viewpoint. Both DRIVE-SMART (e.g. Bruce et al., 2017) and RAPT (e.g. Pradhan et al., 2009) have used this approach. Finally, Horswill et al. (2017) argue that providing advice in the form of feedback (e.g. showing at which point a safe driver would have responded, compared to the trainee's own response), can improve subsequent hazard perception performance. They contrasted their feedback-training results, which were positive, with those of Dogan et al. (2012), who failed to find feedback benefits. They argued that their focus on feedback to individual hazards, rather than the kind of aggregated feedback provided by Dogan et al., was an important factor in this success. RAPT also provides error-based feedback training, with drivers allowed to make and learn from errors in a safe environment (Pradhan et al., 2009).

The current best practice as advocated by US researchers, and the one adopted for this study, is termed ‘3M training’ (e.g. Agrawal et al., 2018; Fisher et al., 2002; Pradhan et al., 2009). It refers to a three-stage process. First, trainees have an opportunity to test their skills in an initial assessment, during which they are likely to make one or more *mistakes* (the first ‘M’). Once they realise that there is room for improvement, they are provided with error-based feedback and are told how to avoid such mistakes in the future. This is termed the *mediation* stage, the second ‘M’. Finally, they demonstrate their improved skills in a similar test environment to stage 1’s. This third ‘M’ allows them to demonstrate *mastery* of the new ability.

The question remains, however, whether these hazard training techniques will be better served in a 360-degree environment. As noted in the introduction, two published studies have reported attempts to do this (Agrawal et al., 2018; Madigan & Romano, 2020). As noted in Chapter 1, Madigan and Romano acknowledged that their study confounded exposure time and presentation mode, reducing the robustness of their beneficial training effects. While the study by Agrawal et al. (2018) was less confounded, their results showed a VR-training benefit only when it came to whether drivers were more likely to fixate subsequent hazards in a simulator. A second behavioural measure (mitigation of the hazard via simulator controls) did not show VR to give better results than single-screen training. As fixation on an object does not necessary accord with correct processing and identification of that object, it is difficult to conclude that VR training was superior without a corresponding benefit in a subsequent behavioural response. Given the equivocal results regarding VR training of hazard awareness, Study 4 was undertaken to help identify any benefits of this training mode.

5.1.2 The current study

The aim of Study 4 was to compare the hazard awareness skills of three groups of drivers following either VR training, single-screen training or a control training condition. Participants first underwent two tests of their hazard perception skills (using a driving simulator, and a video-based hazard prediction test displayed on a single screen). These tests provided baseline measures of performance that could be subsequently co-varied out of post-training performance measures. Participants were then given either VR training, single-screen training or a control training condition (i.e. a driving-related filler task with no anticipated training benefit). The VR and single-screen training procedures were identical apart from the presentation mode. Training consisted of our ten CGI clips. Each clip was presented as a test clip, with participants attempting to predict the hazard following occlusion (allowing them to make a *mistake*). Following a response, the full clip was then replayed with feedback on where and why drivers should have looked in certain areas of the scene. A range of feedback devices were employed including an expert voice-over, overlaid ellipses demonstrating where one should look, and alternative perspectives on the scene (e.g. viewing the scene through the eyes of another road user, or being shown a plan view of the scenario via a virtual satnav). According to Agrawal et al. (2018) this equates to the *mediation* stage (though *mitigation* might be a better term for it). Finally, participants were allowed to demonstrate their *mastery* by undertaking a second simulator drive and a second video-based hazard prediction test in VR.

As noted above, the pre-training and post-training assessment used video-based clips developed in Study 2. The training clips, however, were developed from the CGI clips used in Study 3. The rational for this division was that the CGI clips' relative lack of complexity, and the more-explicit (designed) structure underlying the hazards, would provide a more scaffolded training experience for our drivers. Whereas the rich, cluttered scenes in the naturalistic videos might distract from the learning points (i.e. where to look to detect particular hazards), the CGI clips were thought to allow drivers to focus on the key feedback messages. Conversely, the video-based hazard clips were an ideal choice to measure training outcomes as they reflect the real world more closely. If training in the CGI clips is to have any real-world benefit, it should arguably improve performance on the post-training hazard prediction test.

A further design point to note is that our baseline measure of hazard prediction was obtained from the single-screen version of the Study 2 test, whereas the post-training test was presented in VR. This was done for several reasons. First, we did not want to give drivers in the non-VR-training conditions any exposure to VR prior to the final hazard test, as any such exposure may have confounded the training conditions. Second, we wanted the final measure of hazard prediction skill taken in this study to be as close to the real visual scene as possible. For this reason, the final test was given in VR to all three groups. A final benefit of this design follow from the fact that participants had to be informed (for ethical reasons) that the study involved VR before they volunteered to take part. Many participants came to the study looking forward to trying out the VR test, and the use of the VR assessment at the very end of the session ensured that no participants felt cheated out of this experience.

We predicted that participants who took part in either the VR training or the single-screen training would show differences in behaviour compared to the control group, both on a subsequent simulator drive, and in the video-based VR hazard prediction test (after accounting for baseline performance). Furthermore, we predicted that the VR-training group would show even greater training benefits than the single-screen training group.

5.2 Method

5.2.1 Participants

Ninety-nine participants were recruited across a range of driving experiences, split equally across each of the three training interventions (control, single-screen training and VR training). Of these drivers, 22 were classified as novice drivers (either learning to drive or having less than one year of driving experience) and 77 as experienced drivers, though these were allocated equally across the training interventions. The driver split was not intended as a factor to be analysed in the study, but was merely a reflection of the participants who were available to take part during the data-collection phase (which took place in the latter half of 2020, during which the university was mainly closed, owing to the COVID-19 pandemic).

Three participants were removed owing to sickness encountered during the study (two participants from the single-screen condition and one from the VR condition). It should, however, be noted that the two removed from the single-screen condition were removed as a result of simulator sickness (i.e. symptoms induced by our fixed-base driving simulator), whereas only one participant was removed as a result of cybersickness from exposure to the VR test. The demographic details of the participants in each of conditions, and those participants who suffered sickness, are given in Table 5.1.

Table 5.1: Demographics of all participants in each training condition who completed Study 4, showing participants who suffered from sickness in grey at the bottom of the table

Condition	N	Gender	Mean age (years)	Mean driving experience (years since passing driving test)
Virtual reality – experienced*	25	15 females	37.9	17.6
Virtual reality – novice	7	3 females	21.6	0.1
Single-screen – experienced*	25	14 females	37.9	18.0
Single-screen – novice	7	6 females	20.2	0.1
Control – experienced	24	14 females	38.6	18.4
Control – novice	8	6 females	26.2	0.1
Virtual reality – experienced	*1*	*1 female*	*50.3*	*32.5*
Single-screen – experienced	*2*	*2 females*	*55.6*	*31.4*

Source: Authors' own (Study 4)
Note: Not including participants removed due to sickness

5.2.2 Design

A 1 × 3 between-groups design compared performance on all the dependent variables recorded in the post-test assessments across the three training conditions. Participants were pseudo-randomly allocated to one of the three conditions on the basis of their driving experience (which was taken from their demographics questionnaires prior to coming into the laboratory). This was done to ensure that driving experience did not differ significantly across the three groups. The first condition was *VR training,* in which participants completed the CGI training in a VR headset. The second condition was *single-screen training*, in which participants completed the same training via a standard computer monitor. The third condition was the *control condition,* in which participants had no exposure to the training and completed a filler task instead.

The dependent variables included post-training measures of performance in the simulator and performance on a post-training hazard prediction task. The main dependent variable from the simulator was the number of collisions that each driver had, though more detailed data regarding average speed, speed variance, mean lateral position, mean wheel error (how much the angle of the steering wheel deviated from the most consistent path)and lateral position variation was also collected.

The dependent variable from the post-training hazard prediction test was the number of hazards correctly predicted. Pre-test performance on the driving simulator and a single-screen hazard test were used as covariates. We predicted that all trained drivers would have fewer crashes in the simulator and higher scores on the post-training hazard prediction test than control participants. We further anticipated that the training effects in the VR condition would be superior to those of the single-screen training condition.

5.2.3 The training conditions

5.2.3.1 The VR and single-screen training interventions

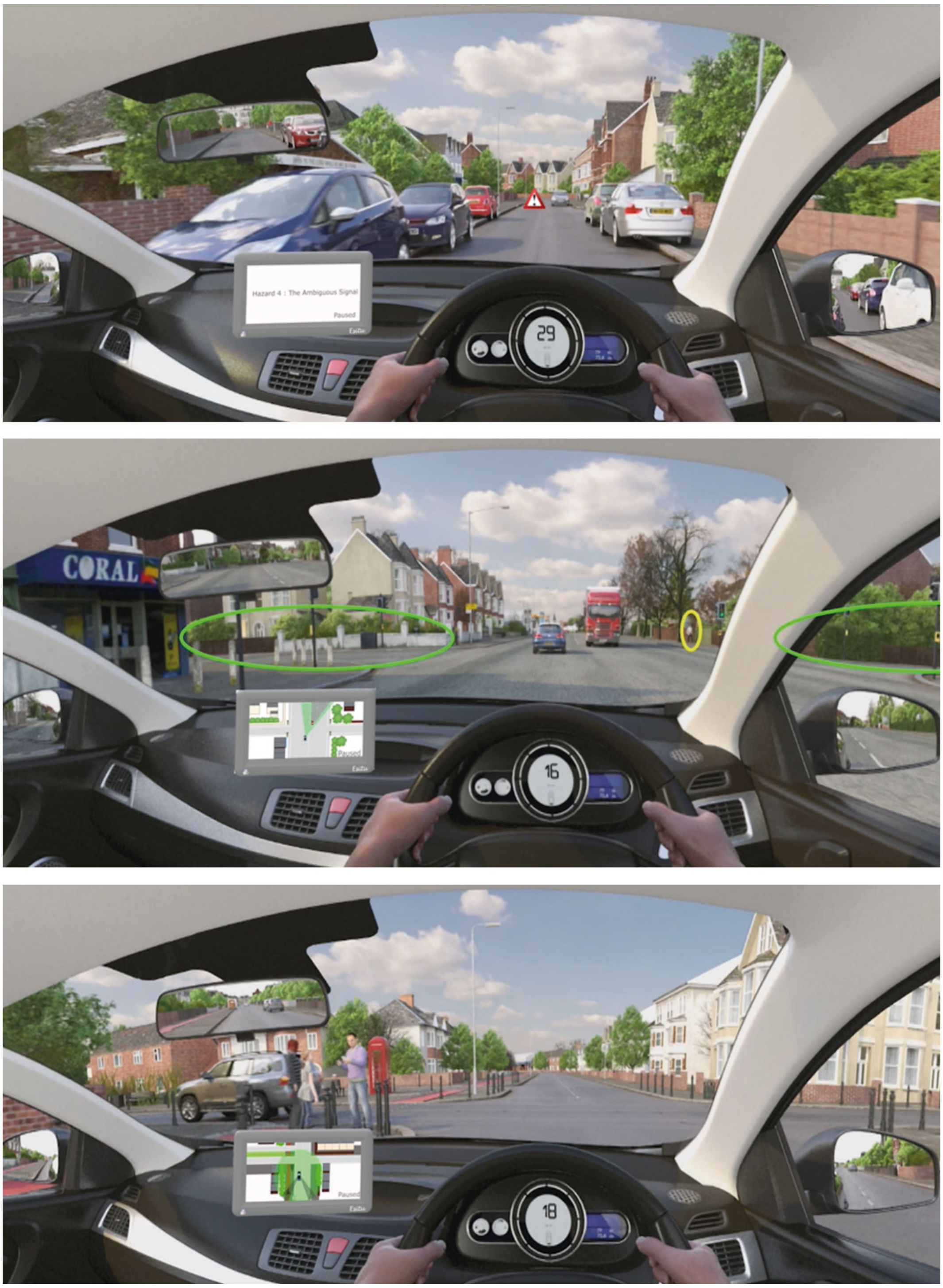
The VR-training condition and the single-screen training condition were identical except for their mode of presentation. The training was based on the ten CGI clips used in Study 3. Participants would first view a clip under test conditions and try to select the correct answer from the four options presented post-occlusion. Following a response, the clip would be played again in full, with an expert voice-over advising on where to look and why. The voice-over was accompanied by graphical overlays (e.g. coloured ellipses) to point out areas of the scene that should be attended to. Other graphics included a warning sign to reinforce the hazardous nature of the situation and an arrow to show the projected trajectory of another road user. Further elements of feedback included providing the perspective of another road user (e.g. the car driver who pulls out of a side road in Hazard 4: see Table 4.2), and top-down schematic views provided by a virtual satnav that was visible on the dashboard of the film car in the replayed clips. During particularly complicated sections of clips, the playback speed would be slowed down to allow the voice-over to point out all relevant information. Where additional reinforcement was required, the clip would even rewind to play important sections again. Screenshots containing examples of the visual aids used to provide feedback training can be viewed in Figure 5.1. A close-up of the virtual satnav is provided in Figure 5.2.

5.2.3.2 The mind-wandering filler task

The filler task was a ‘mind-wandering task’ that requires participants to watch a ten-minute clip of predominantly rural and dual carriageway driving. The format of the test was designed to be similar to that of the hazard prediction test (presenting mirror information within a graphic overlay of a car interior), though nothing of interest happened in the clip. Participants were asked to watch the clip as if they were the driver. Periodically, the clip paused, and drivers were asked to rate how focused they were on the driving task on a 1–7 scale. Participants were made aware in advance that these probes would occur during the test.

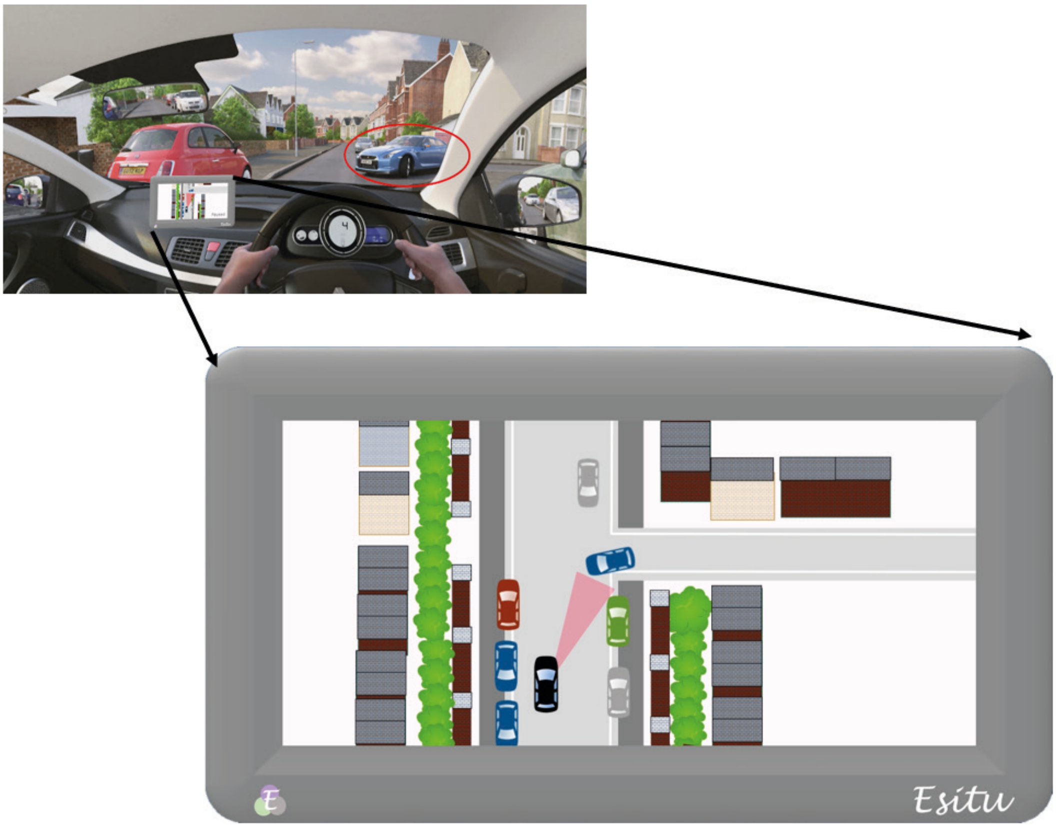
Although driving-related, this task was not anticipated to provide any learning benefit in the post-training assessments, and mind-wandering ratings were not analysed. The test simply ensured that all driver groups were exposed to the same amount of driving stimuli.

Figure 5.1: Three screenshots showing different forms of visual feedback in Study 4: a warning symbol (top panel), the use of ellipses to indicate where drivers should look (middle panel), and the use of a virtual satnav to provide alternative perspectives on hazards (middle panel and bottom panel)



Source: Authors' own (Study 4)

Figure 5.2: Close-up of the virtual satnav used in the CGI training clips to provide top-down perspectives on hazards in Study 4



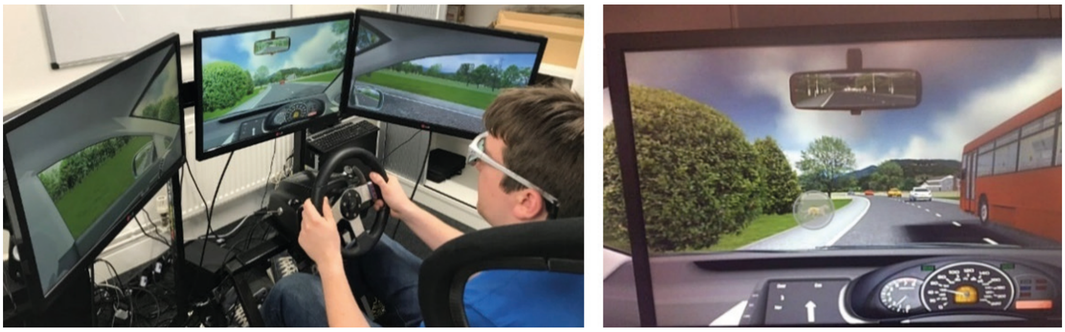
Source: Authors' own (Study 4)

5.2.4 The tests of training benefit

5.2.4.1 The driving simulator assessment

In a previous project funded by the Road Safety Trust, which examined the impact of a mindfulness intervention on driver safety (Crundall, Kroll, Goodge and Grifffiths, 2019), two simulated routes were programmed and subsequently validated for use on a Carnetsoft driving simulator (see Figure 5.3). Each route contained ten hazards (see Table 5.2 for a description of the hazards that appear in the simulator). The hazards were matched across both routes in terms of the type of danger they represented. Participants completed one drive in the pre-training assessment and one drive in post-training assessment. The order of the routes was counterbalanced across participants (e.g. half of the participants underwent Route A in the pre-training assessment and Route B in the post-training assessment, while this order was reversed for the other half of the participants). Measures such as speed, steering wheel angle, and frequency of collisions were calculated as dependent variables to assess the impact of the training interventions.

Figure 5.3: Three-screen Carnetsoft simulator (left panel) and screenshot of the central screen as Hazard 1 (Table 5.2) triggers (right panel)



Source: Crundall et al., (2019)

5.2.4.2 The hazard prediction tests

Twenty clips were selected from Study 2 to form the basis of two hazard prediction tests for the current study. Both tests were prepared in VR and single-screen format, and clips were matched and assigned to clip sets on the basis of Study 2 data, to provide tests of similar difficulty. The clips assigned to clip set A were 1, 2, 3, 5, 7, 11, 13, 16, 18 and 23. The clips assigned to clip set B were 4, 6, 8, 10, 14, 15, 19, 20, 22, 24 (see Table 3.2).

5.2.5 Apparatus

Three of the tasks (the single-screen hazard prediction test, the single-screen training intervention, and the mind-wandering task) were run from a Lenovo ThinkPad laptop connected to a large monitor. The monitor measured 60 × 34 cm and was positioned at approximately 60 cm from the participant, creating a visual angle of 53 (horizontal) by 32 (vertical) degrees. Participants listened to the voice-over in the single-screen hazard prediction test and the expert commentary on the training videos via speakers that were attached to the laptop.

Both the VR hazard prediction test and VR training were run on an Oculus Go VR headset which was connected via a cable to a Lenovo laptop, so that the researcher could view the participants' progress on the laptop screen. In both tests, participants gave their answer verbally to the researcher, who recorded their answer. Participants were able to hear the voice-over and commentary through the speakers in the headset.

To sterilise the VR headset between participants, a UVC (ultraviolet C) light box (Cleanbox) was used. The medical-grade UVC light is reported to kill 99.9% of bacteria, viruses (including COVID-19) and fungi within 60 seconds.

The driving simulator task was completed on a Carnetsoft driving simulator. The Carnetsoft simulator consists of three screens, a bucket seat, a steering wheel, a gear stick and a pedal set. The seat is adjustable to ensure that all participants can reach the pedals. Participants drove the simulator as a manual vehicle and followed auditory instructions produced by the simulator while navigating the route.

Table 5.2: Description of the ten pairs of matched hazards that participants would encounter in the driving simulator created for Study 4 (Drive A and B)

Hazard number	Order presented (Drive A)	Drive A	Order Presented (Drive B)	Drive B
1	1st	A dog emerges from behind a bush on the left and runs into the road.	3rd	A dog runs into the road from the right.
2	10th	A stopped lorry on the left indicates and pulls out to join the road.	2nd	A bus stopped at a bus stop indicates and pulls out to join the road.
3	8th	A pedestrian emerges from behind a stopped bus and crosses the road in front of the film car.	6th	A pedestrian emerges from behind a parked lorry with its hazard lights flashing and crosses the road in front of the film car.
4	4th	A parked vehicle pulls out from a lay-by and then brakes suddenly to avoid a dog that runs into the road.	8th	The vehicle ahead brakes suddenly and then comes to a stop and turns its hazard lights on.
5	2nd	A parked vehicle in a lay-by to the left indicates and pulls out to join the road.	7th	A parked vehicle in a lay-by to the left indicates and pulls out to join the road (a different vehicle to that in hazard 5).
6	3rd	A pedestrian steps onto the zebra crossing from the right.	4th	A pedestrian steps onto the zebra crossing from the left.
7	5th	An oncoming vehicle pulls onto the film car's side of the road to overtake an oncoming bus.	10th	An oncoming vehicle pulls onto the film car's side of the road to overtake an oncoming lorry.
8	7th	A vehicle waiting in a side road on the right pulls out in front of the film car.	5th	A vehicle waiting in a side road on the left pulls out in front of the film car.
9	6th	An oncoming vehicle cuts across the film car as it turns into a side road on the left.	9th	An oncoming vehicle cuts across the film car as it turns into a side road on the left (a different vehicle to that in hazard 9).
10	9th	A parked vehicle encroaches on the road, requiring the film car to manoeuvre around it.	1st	A parked vehicle encroaches on the road, requiring the film car to manoeuvre around it.

Source: Authors' own (Study 4)

5.2.6 COVID-19 protocol

The testing phase for Study 4 fell during the coronavirus pandemic. Initially all data collection was halted across the university. Working with the Nottingham Trent University Health and Safety team, we created a COVID-19 protocol for interacting with participants during a pandemic. This consisted of numerous precautions including participant greeting procedures, equipment sterilisation, the use of face coverings by both researcher and participant, and the use of nitrile gloves. As already mentioned, a UVC light box was purchased to sterilise the VR headset, while all other equipment was cleaned with alcohol wipes between participants. Rules on participant recruitment were tied to whichever COVID-19 local restriction tier the university fell under at that point in time.

The COVID-19 protocol was accepted by the university, along with a revised risk assessment and ethics application. Testing was allowed to commence in September 2020, eventually ending with the most recent lockdown (December 2020). Our COVID-19 protocol was subsequently adopted by the British Psychology Society and was put on their website as guidance for other psychology departments who sought to return to the laboratory.

5.2.7 Procedure

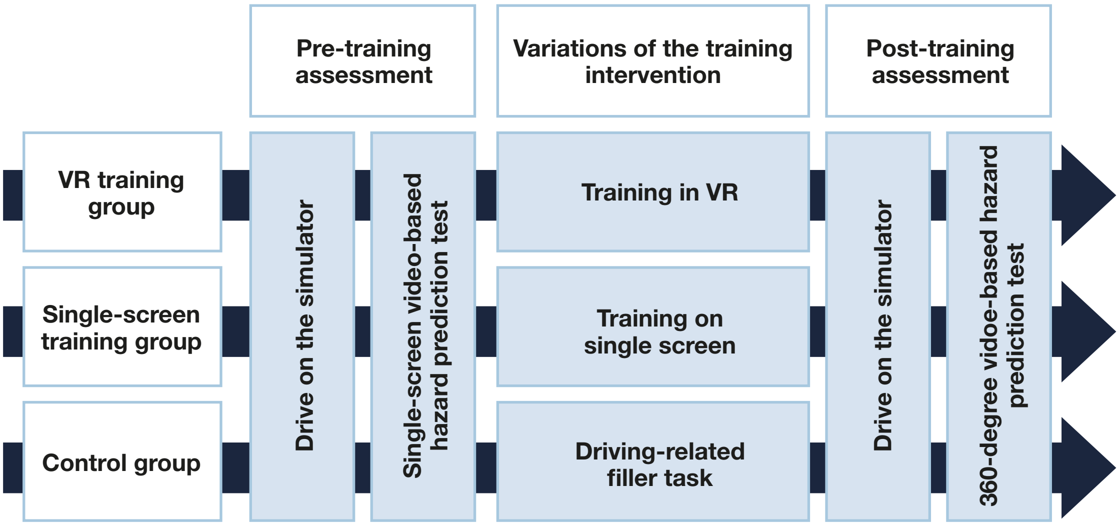
Immediately after signing up, participants were sent a link to a demographics questionnaire via Qualtrics which they were asked to complete prior to their laboratory session. Participants were also advised that we would be following a strict COVID-19 protocol, including the use of face coverings, hand sanitiser, track and trace, and social distancing wherever possible. They were sent a digital information sheet and consent form and, after having had the opportunity to ask questions, were asked to complete the consent form on their mobile phone. Participants then completed all the tasks for their respective condition, which took between 1.5 and 2 hours per participant.

Prior to training, all participants received a five-minute practice drive on the simulator before completing the first of the simulated hazard routes. A simulator sickness check was carried out before the practice drive, after the practice drive and after the initial drive. The SSQ was reduced to a single item querying participants' general sickness symptoms on a scale of 1–20. This reduced SSQ was used to lessen the time required for participants to be in the laboratory, balancing the greater validity of the full SSQ against the increased risk of COVID-19 transmission by keeping participants in the laboratory for longer than necessary. Following the simulator test, participants undertook a video-based hazard prediction test presented on a single screen.

Participants were then assigned to one of the training conditions: VR training in the Oculus Go headset, single-screen training, or the mind-wandering filler task. For those participants in the VR-training condition, their sickness symptoms were monitored at multiple points using the single-item SSQ.

Following the training intervention, participants undertook a second simulator drive, and a video-based VR hazard prediction test. All participants had their sickness symptoms monitored throughout the final VR test. The order of the tests remained the same across all participants (see Figure 5.4), though the simulated driving routes, and hazard prediction clip sets were counterbalanced. All participants received a £30 online voucher for taking part in the study.

Figure 5.4: Schematic depiction of the procedure of Study 4


Pre-training assessment
Variations of the training intervention
Post-training assessment
VR training group
Single-screen training group
Control group
Drive on the simulator
Single-screen video-based hazard prediction test
Training in VR
Training on single screen
Driving-related filler task
Drive on the simulator
360-degree vidoe-based hazard prediction test

Source: Authors' own (Study 4)

5.3 Results

For all analyses, a cut-off of two standard deviations was used for identifying outliers. This cut-off was calculated from the whole cohort (across training condition and experience groups) from their pre-test scores. For all group and training comparisons we computed analyses of covariance (ANCOVAs). These analyses compare the pre-training measures (e.g. simulator performance, hazard prediction score) according to which training intervention they had undergone, while co-varying the pre-intervention measures. Essentially, this allowed us to compare participants' scores in the lab following the training courses, while statistically accounting for natural variation in baseline performance.

5.3.1 Hazard prediction performance

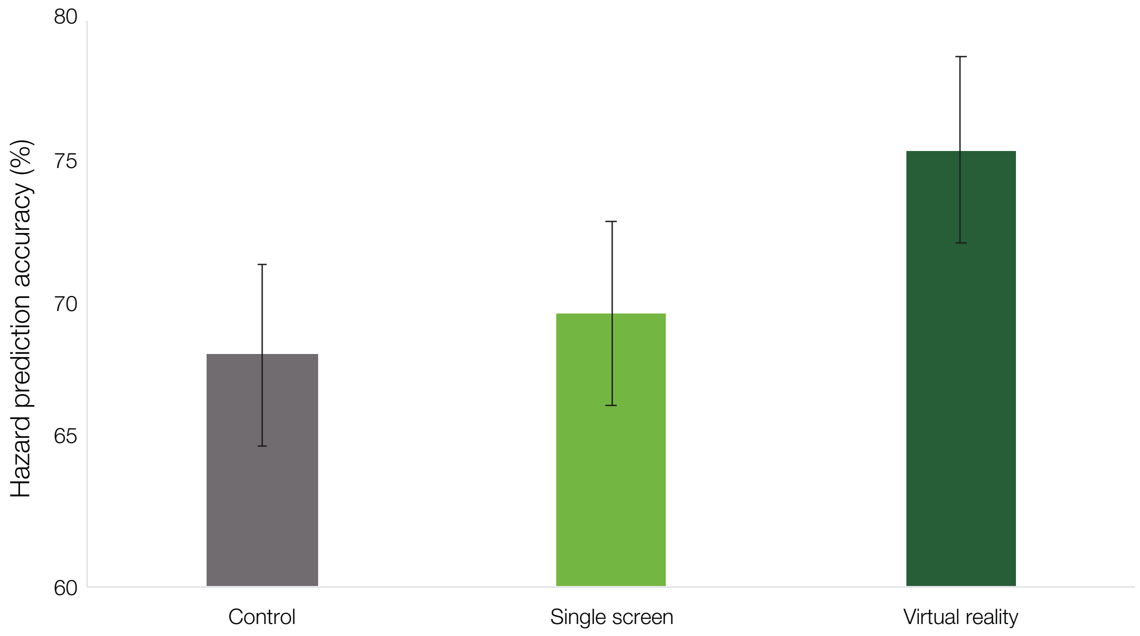
One participant was removed owing to missing post-test data. A further six participants were removed as outliers because their pre-test scores fell +/–2 standard deviations or more beyond the sample mean. This suggests that these participants were not engaged with the task or were unsure of the task procedure.

Participants' percentage accuracy rates for correctly predicting the hazards during the post-test hazard prediction test were compared across the training groups in a one-way between-subjects ANCOVA, with their pre-test scores as covariate. As can be seen in Figure 5.5, there is an apparent trend for participants who received VR training to be more accurate (*M* = 75%) than either the participants trained in the single-screen condition (*M* = 70%) or the control participants (*M* = 68%). However, this difference did not reach the threshold for significance ($F(2, 85) = 1.4$, $MSE = 313.6$, $p = .26$, $\eta^2 = .03$).

Whilst the means of the groups follow the predicted pattern, there are a number of potential reasons why this difference did not reach significance. One possibility is that the statistical power of the study was not sufficient. However, power analyses are based on predicted effect sizes noted in similar studies, such as that by Agrawal et al. (2018). In that study, the researchers found a VR-specific training benefit with the same number of assessment scenarios, and approximately one third of our sample size. Furthermore, their overall training effect sizes (regardless of presentation mode) were extremely high. Using Agrawal et al. to calculate necessary sample sizes suggests that we should have found an effect if there was one there to be found.

One issue that may have impacted our power, however, is the heterogeneity of our sample. Our initial intention for Study 4 was to test only novice drivers. This high-risk group is our primary target audience, and these drivers are likely to have the greatest capacity for value to be added to their skill set through training. Unfortunately, COVID-19 restrictions affected our recruitment strategy (effectively removing our student population), resulting in a wider range of participants. This may have increased the variance in performance across the group. Accordingly, the data was also modelled using a multilevel binomial logistic regression with participants and clips as random factors, and training as a between-groups fixed effect. This analysis should have reduced the impact of individual variation, but still did not reveal a significant effect.

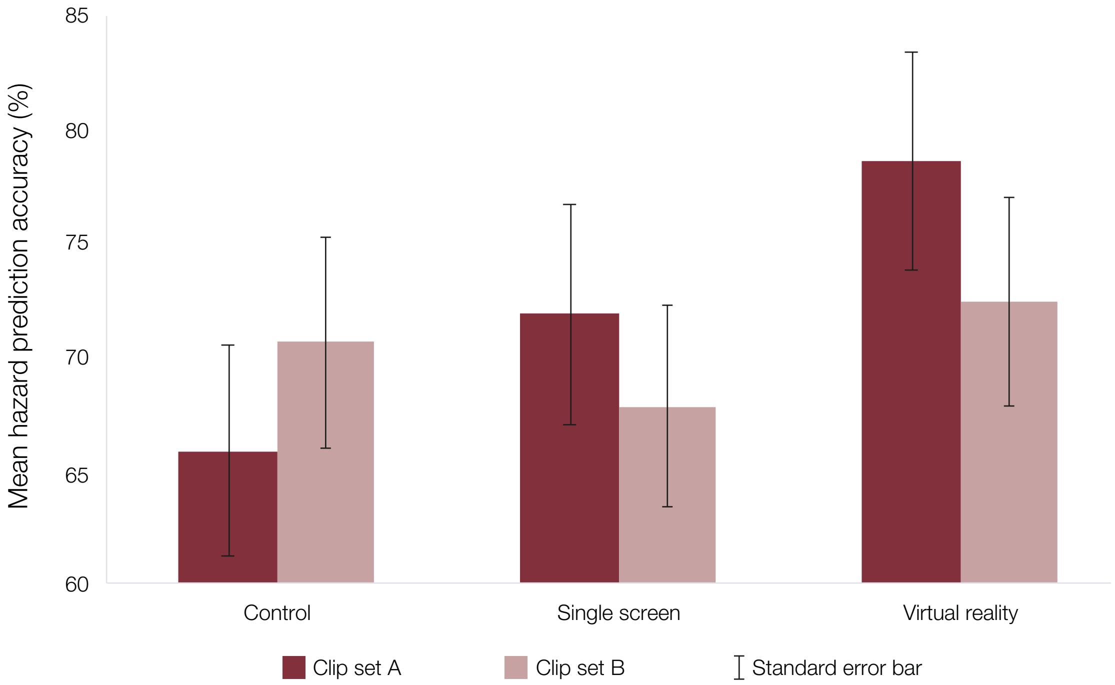
Figure 5.5: Average hazard prediction accuracy across the training groups in Study 4



Source: Authors' own (Study 4)

A further possible explanation is that the training improves performance on one of the clip sets more than the other. Although the clip sets were matched for difficulty, it is possible that one set contains precursors that benefit more from the training (i.e. clips that have underlying similarities to training scenarios and thus allow 'near transfer' of training). Although the interaction was not significant (p = .46), Figure 5.6 suggests that, if a training effect were to be found, it would most likely be found with clip set A.

Figure 5.6: Hazard prediction accuracy for post-test clip sets A and B used in Study 4 whilst co-varying pre-test scores



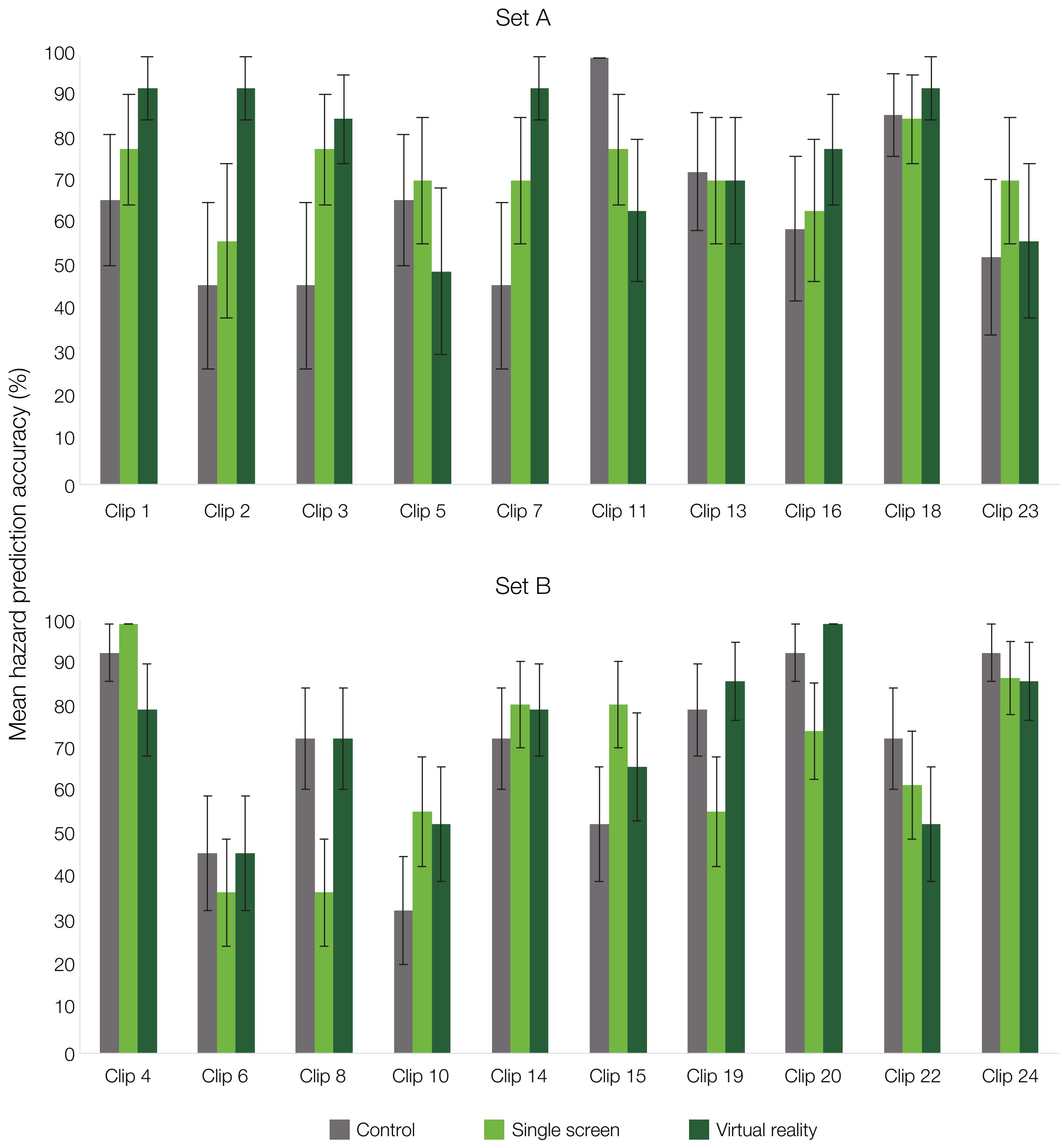
Source: Authors' own (Study 4)

To examine this further, a one-way between-subjects ANCOVA was conducted on post-test accuracy of only set A clips, whilst co-varying pre-test scores. The omnibus calculation was not significant, $F(2, 39) = 2.1$, $MSE = 283.3$, $p = .14$, $\eta^2 = .10$, though planned Helmert contrasts revealed marginal evidence for a difference between VR-trained and control participants, suggesting that the VR-trained drivers significantly predicted more hazards than the control drivers (77.8% vs 64.7%, $p = .05$).

This suggests that the clips in clip set A are closer to showing a training benefit for VR. Drilling down further, we investigated the percentage of correct responses given to each clip within their respective clip sets. For post-test set B (Figure 5.7, bottom panel), only clip 8 suggests a possible training effect for the single-screen training condition. However, when subjected to Fisher exact tests, none of the clips showed a difference in hazard prediction performance between any of the training groups (all values of $p > 0.05$).

For clip set A, however (Figure 5.7, top panel), several individual clips (clips 2, 3, 7 and 11) suggest a possible difference in responses across training groups. When subjected to Fisher exact tests, the group difference between clip 2 and clip 7 showed that VR-trained drivers more accurately predicted the hazards than the single-screen and control participants (both values of $p < 0.05$). The apparent difference in response accuracy to clip 3 did not reach threshold ($p = .08$), while clip 11 oddly suggests that the training conditions might have decreased performance on this clip compared to the control condition ($p = .04$).

Figure 5.7: Hazard prediction accuracy scores across all clips for post-test clip set A (top panel) and B (bottom panel) across all intervention groups in Study 4



Source: Authors' own (Study 4)

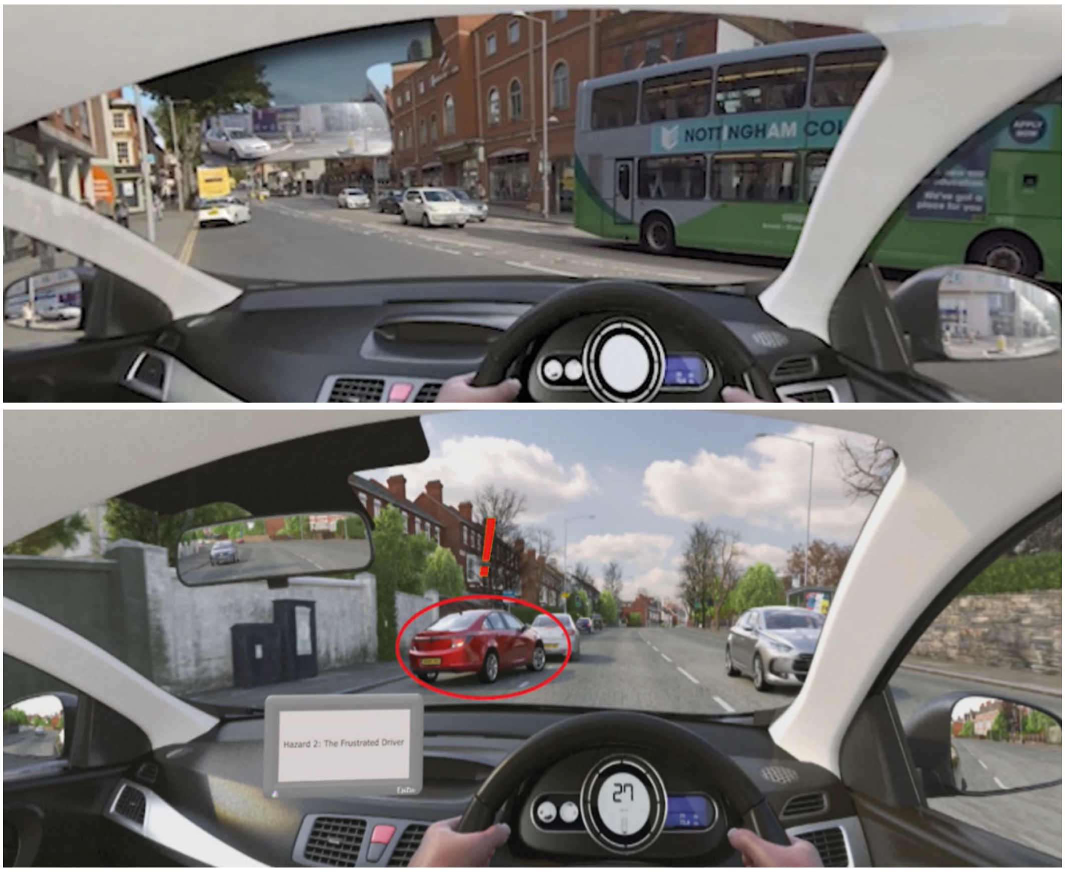
5.3.2 Training benefits for individual clips

As noted in the previous section, four clips from post-training clip set A demonstrated an apparent difference in responses owing to training condition (clips 2, 3, 7, 11; though only three of these clips reached the threshold of significance). This section compares the content of each of these four clips with the training material, in an effort to identify possible routes for 'near transfer' of training.

> *Clip 2: "After you have waited at a set of traffic lights, they turn green and you take a left turn. As you follow the road, a white taxi does a U-turn in the middle of your lane, blocking your path."*

This clip requires a rapid movement of the eyes, a quick stabilisation of the taxi on the participant's retinas and immediate processing of the threat before the sudden occlusion. The taxi is visible only after the film car navigates a bend in the road. There is little time to identify possible reasons that would prompt a hazard, as the time gap between rounding the corner and the occlusion is very short. Thus successful identification of the hazard in this clip requires the viewer to position their eyes at the relevant location (which is the immediate heading of the film car once around the bend) and then rapidly extract information that the taxi is turning in the road (Figure 5.8, top panel). Several training clips involve turning corners, and the requirement to fixate objects in the driver's new pathway. It is possible that our training has encouraged drivers to move their eyes to more appropriate locations in the road ahead as they navigate a corner. Furthermore, it is understandable why such training might be more beneficial when presented in a VR headset rather than on a single screen, as the 360-degree view allows drivers to practice more realistic head and eye movements as the film car enters a new road.

Figure 5.8: Identification of the hazard in the post-training assessment clip (a white taxi preparing for a U-turn: post-training assessment clip 2) may have been primed by exposure to similar behaviours of road users in the CGI training clips (training clip 2, bottom panel)



Source: Authors' own (Study 4)

A second possibility is that the training sensitised participants to the image of a car at an angle. The image of a car ahead, oriented to an angle that is not concordant with the direction of the road, is highly salient. This finding has led to the practice of police stopping at an angle on the hard shoulder of motorways (Langham et al., 2002). This stopping method was argued to reduce motorway collisions with stationary police cars as the transverse position makes it clear to other drivers that the police car is stationary rather than travelling. Several training clips contain other vehicles moving at angles contrary to the direction of the road, suggesting an imminent hazard. For instance, training clip 2 (Figure 5.8, bottom panel) shows a car in the left lane pulling into the film car's lane to make an overtaking manoeuvre. The training imagery is very close to that of the hazard in the assessment clip (though the reasons for the hazardous manoeuvre differ between the two clips). It is thus possible that this clip drew attention to an already salient cue, and gave it meaning within a hazardous context.

> *Clip 3: "While you are driving along a suburban route, an oncoming police car becomes visible in the distance, and so you have to pull over to give way."*

For clip 3, there was a non-significant tendency for VR and single-screen trained participants to better predict the appearance of an oncoming emergency vehicle than control drivers. Interestingly, training clip 6 did include traffic stopping to let an oncoming police car through. The police car was not considered to be the hazard in the training clip. Instead, it was used as a device to allow a previously observed cyclist catch up with the film car. The cyclist subsequently becomes the hazard. Nonetheless, it is possible that the training clip sensitised drivers to stay alert for emergency vehicles, thus better preparing drivers for the subsequent assessment clip (Figure 5.9). This may have occurred by increasing the cognitive salience of flashing lights in the distance, or by linking the behaviour of other road users (who pull over in both the training clip and the assessment clip) to the subsequent appearance of an emergency vehicle.

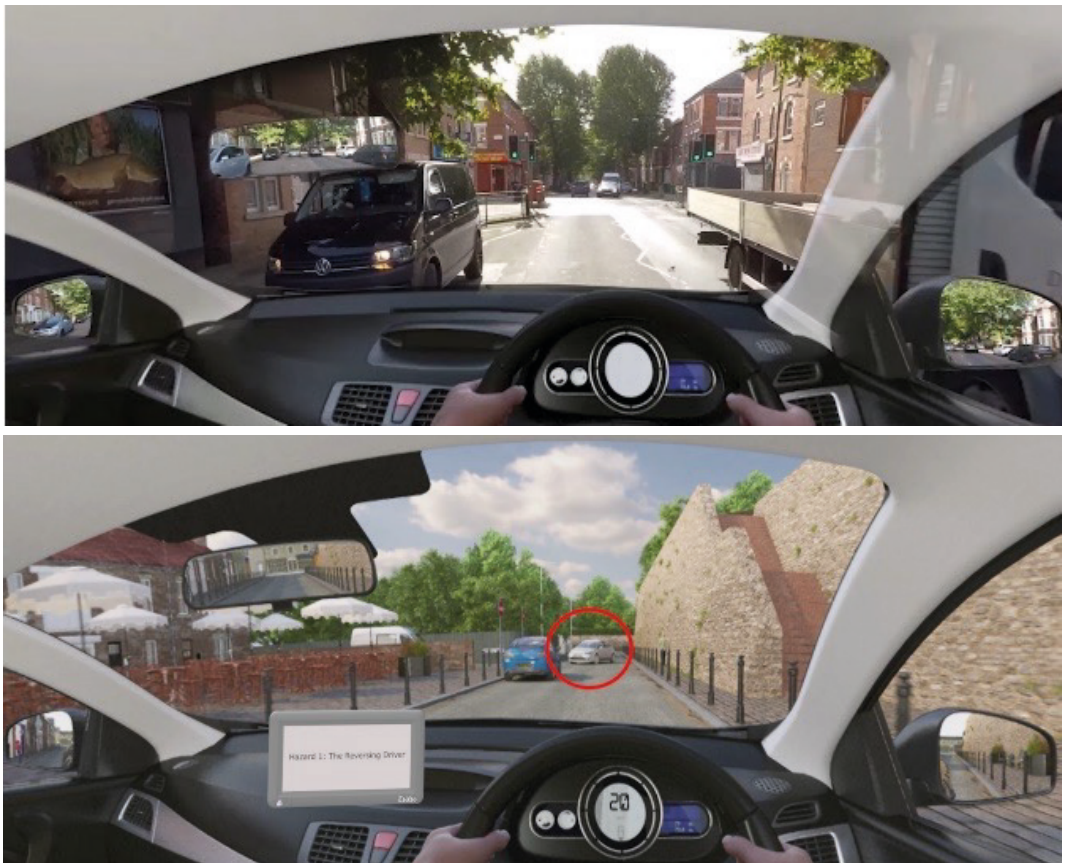
Figure 5.9: Response to the hazard in the post-training assessment clip (the blue flashing lights of an oncoming police car are visible in the distance: post-training assessment clip 3) may have been primed by exposure to a similar scenario in the CGI training clips (training clip 6, bottom panel)

Source: Authors' own (Study 4)

Clip 7: "As you are travelling along a busy urban route, an oncoming car pulls across your path into the side road on the left, after a set of traffic lights."

This post-training assessment clip requires the driver to look far down the road, where they will notice that an oncoming car has stopped and begins to turn across their path (Figure 5.10, top panel). The behaviour of this other road user is mirrored in training clip 1, in which an oncoming car performs the same manoeuvre. In the training clip, the oncoming car should be far enough away to make the turn into the side road before the film car reaches that location (Figure 5.10, bottom panel). The turning car is, however, blocked from entering the side road by a van, and thus blocks the path of the film car. This training clip may have emphasised the need to look for turning vehicles further down the road just in case their intended manoeuvre is not successfully completed by the time one arrives at that location. The VR training may have been especially helpful in this clip, as the larger scene would have allowed the viewers to look further down the road than on a single-screen clip.

Figure 5.10: Response to the hazard in the post-training assessment clip (the oncoming car turns across your path: post-training assessment clip 7) may have been primed by exposure to a similar scenario in the CGI training clips (training clip 1, bottom panel)

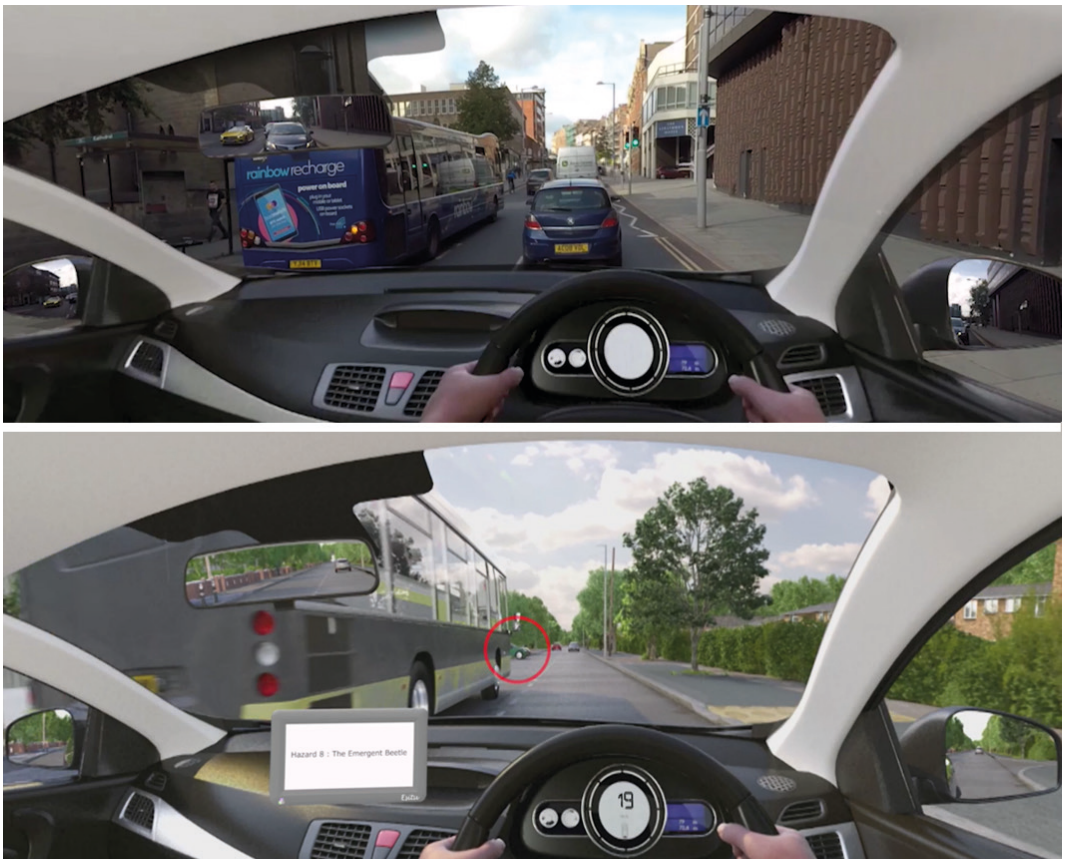


Source: Authors' own (Study 4)

Clip 11 – "While you are driving in congested traffic, a bus on your left indicates to pull into your lane to overtake a cyclist."

Interestingly, for clip 11, the control drivers outperformed both the VR and single screen trained drivers. The hazard in this clip involves a bus in the left-hand lane, indicating to pull into the right-hand lane in front of the film car. The clip occludes at the moment the bus begins its manoeuvre (see Figure 5.11). A potential explanation of this finding is that in both the VR and single-screen training conditions, participants are trained to spot subtle cues in the environment to predict hazards. This may have inadvertently led them away from larger, more obvious hazards. A second possibility is apparent in training clip 8. This clip contains a static bus, from behind which a car emerges (Figure 5.11, bottom panel). It is possible that, when faced with a nearby bus, our trained participants view the bus as a potential obscurer of hazards, rather than a hazard in itself.

Figure 5.11: Response to the hazard in the post-training assessment clip (the bus pulls into your lane: post-training assessment clip 11) may have been ignored owing to a similar training hazard that primed a different hazardous outcome (training clip 8, bottom panel)



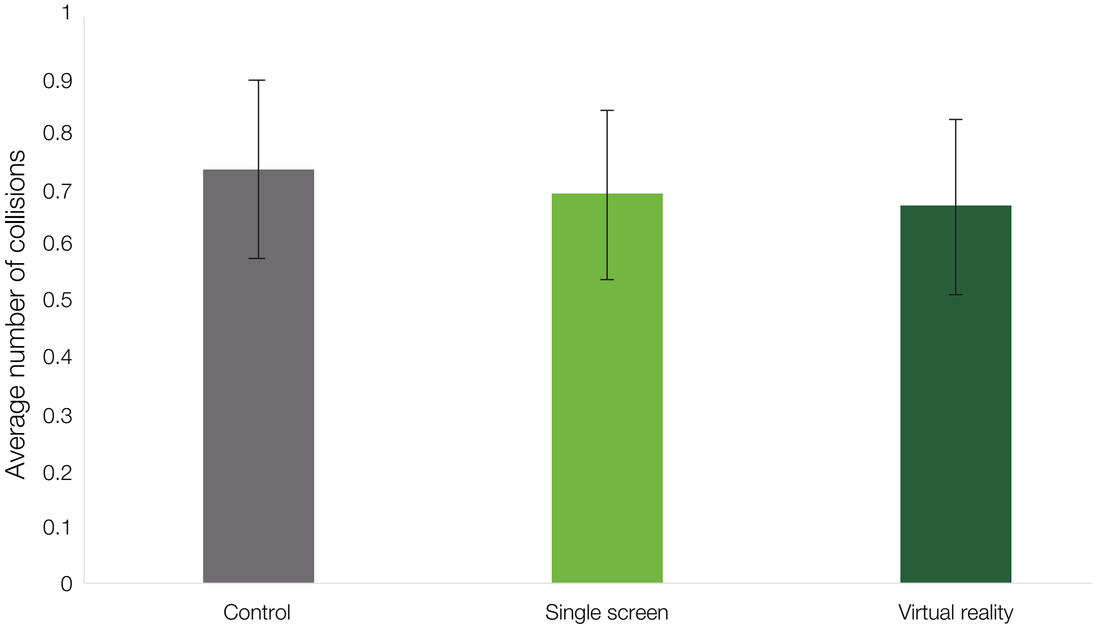
Source: Authors' own (Study 4)

5.3.3 Driving simulator performance

Seven participants were removed from all simulator data analyses. Four of these participants were removed because of data loss (one control participant, two VR-trained participants and one single-screen trained participant). The remaining three participants were removed as their number of pre-test collisions were two standard deviations above the group mean of the pre-test collisions (one control participant and two single-screen participants). This suggests that they had trouble with the simulator interface rather than being a bad driver per se.

The number of collisions during the post-training simulator assessment was compared across the three training groups using a one-way between-subjects ANCOVA, with pre-test collisions as a covariate. As can be seen from Figure 5.12, although the average number of collisions goes in the predicted direction, this difference did not reach significance, *F*(2, 85) = 0.1, *MSE* = 0.7, *p* = .95, η^2 = .001.

Figure 5.12: Average number of collisions across the groups



Source: Authors' own (Study 4)

Though the design of the study was not intended to compare novice and experienced drivers, the number of novices was sufficient to allow a comparison, albeit comparing 69 experienced drivers with 20 novice drivers. As above, the number of collisions during the post-training simulator assessment was compared for novice and experienced drivers using a one-way between-subjects ANCOVA, with pre-test collisions as a covariate. When the means were adjusted for the covariates, experienced drivers were found to have 0.6 crashes while novices had 1.1 crashes. This significant difference, *F*(1, 86) = 6.4, *MSE* = 0.7, *p* = .01, η^2 = .07), validates the simulated scenarios, demonstrating that they tap into skill differences between the groups.

A selection of behavioural measures of driver performance were also collected from the simulator (speed, speed variance, lateral position, lateral position variation and mean wheel error). These were calculated across the whole drive for each participant. All these measures were also subjected to the same one-way between-subjects ANCOVA as reported above, with pre-test measures as the covariate. For overall speed, although the omnibus calculation did not reach accepted levels of significance, $F(2, 85) = 2.5$, $MSE = 7.7$, $p = .09$, $\eta^2 = .06$, planned Helmert contrasts revealed a significant difference between VR-trained and control participants, suggesting that the VR-trained drivers chose to drive at significantly slower speeds than the control drivers (with mean speeds across the whole route of 23.7 mph and 25.1 mph, respectively; $p = .03$). None of the other measures reached the threshold for significance (all values of $p > 0.05$).

In order to examine whether participants' driving performance varied when encountering the hazards, we calculated their driving performance for a five-second window prior to the hazard being triggered and compared it with a five-second window prior to a hazard collision (or virtual collision point if the participant did not crash). All of the same measures as used above were subjected to a 2 × 3 ANCOVA across time window (non-hazard window vs hazard window) and training group, with pre-training simulator performance in both windows used as covariates.

For mean speed, there was a main effect of timing, $F(1, 84) = 8.7$, $MSE = 4.9$, $p = .004$, $\eta^2 = .09$, demonstrating that all drivers were on average travelling considerably more slowly in the five seconds before a collision (or virtual collision if they avoided the hazard) than in the non-hazard window (22.7 mph vs 32.5 mph, respectively). Although the omnibus calculation for the training groups did not reach accepted levels of significance, $F(2, 84) = 2.2$, $MSE = 25.6$, $p = .12$, $\eta^2 = .05$, planned Helmert contrasts revealed a marginal significant difference between VR-trained and control participants, suggesting that the VR-trained drivers were slower than the control drivers (26.9 mph vs 28.7 mph, $p = .05$). None of the other main effects or interactions of interest reached the threshold for significance.

In the analysis of mean wheel error, the omnibus calculation for a group difference approached significance ($F(2, 84) = 1.2$, $MSE = .5$, $p = .08$, $\eta^2 = .06$), though planned Helmert contrasts revealed a significant difference between VR-trained and control participants, which suggested that the VR-trained drivers were less likely to deviate from the normative wheel angle (given an ideal path) than the control participants (0.94 vs 1.32, $p = .03$). None of the other main effects or interactions of interest reached the threshold for significance.

For variation in lateral position, there was a main effect of group, $F(2, 84) = 4.6$, $MSE = 0.01$, $p = .01$, $\eta^2 = .10$. Post hoc corrected pairwise comparisons revealed that both the VR and the single-screen trained participants were significantly less variable in their lateral position than the control participants (VR = 9.17, single-screen = 0.18, control = 0.24; all values of $p < .05$). This possibly reflects drivers' improved ability to anticipate the need to change lateral position (spotting a hazard earlier allows a driver to make a smaller adjustment to the lateral position, while still avoiding the danger). None of the other main effects of interactions of interest reached the threshold for significance.

5.4 Discussion

The aim of Study 4 was to examine the impact of driver training in a VR environment on subsequent driving simulator and hazard prediction performance. The results provide mixed evidence.

Regarding the impact of training on subsequent hazard prediction scores, there was no overall effect. When analyses drilled down to individual clips, some clips were noted to show a benefit from VR training, and these clips had clear similarities with content in certain training clips. This may be an example of the ability to find ‘near transfer’ of training (training benefit that applies to situations which closely resemble the training scenarios), but not ‘far transfer’ of training (where trainees extract principles from the training that can be applied to more varied real-world situations).

The results of the simulator analyses were slightly less equivocal. One clear finding was that both trained groups of drivers had less variation in their lateral position (i.e. they were better able to keep in a consistent lane position) than the untrained control group. This was supported by pre-planned group comparisons on mean wheel error (a measure of lateral variation). This measure suggested that VR-trained drivers were the most accurate in terms of steering through the route.

Though we did not find an omnibus group effect for speed, the comparisons again suggest that the VR-trained drivers were more likely to drive at slower speeds. Taking this together with the lateral position data, we infer that the VR-trained drivers (and to a lesser extent, the single-screen trained drivers) did not have to make as many course corrections in response to changes in the roadway and other road users’ behaviour (creating hazards, for example), perhaps in part because they were driving more slowly.

Why might hazard perception training result in lower overall speed? One argument may be linked to risk allostasis theory (Fuller, 2011), which suggests that drivers try to maintain task demands and the associated level of risk within an acceptable tolerance range, which can differ from person to person, and also vary for the same individual over time. If the task is too hard and the risk too great, a driver will attempt to reduce task demands and moderate the danger. Equally, however, if a task is too easy, drivers will be motivated to increase task difficulty and the associated risk to maintain their interest.

This process is prone to error, however, if a driver cannot correctly estimate the level of risk that they are taking. A driver with poor hazard perception skill may believe the road ahead to be devoid of danger for the foreseeable future and therefore decide to increase risk-taking behaviour to maintain task engagement. However, if they have misread the road, the task demands may change so quickly that the driver does not have time to react. If hazard perception training gives drivers a better understanding of the myriad cues that might precipitate a hazard, this may increase the level of perceived risk, with a concomitant reduction in drivers’ willingness to engage in risky behaviours. This was recently noted by Krishnan et al., (2019). They found drivers who received hazard perception training were less likely to engage in secondary activities, such as using a mobile phone, during a simulated drive.

The positive effects of training on simulator behaviour did not, however, result in fewer crashes in the VR-trained group. This arose possibly for the same reasons that plague large-scale naturalistic studies of crash risk (e.g. the Ipsos MORI evaluation of the National Speed Awareness Course, 2018), such as the relative infrequency of crash events, and the multiple causes of collisions that might not be the fault of the trained driver.

5.4.1 Why such equivocal findings?

One factor to consider when assessing marginal results is whether the design had sufficient statistical power to identify the effects. Power analyses and effect sizes from the limited literature suggest that our target N should have been sufficient, especially as Agrawal et al. (2018) were able to find VR-training benefits with approximately one third of the sample size in the current study. Unfortunately, other factors may have had an impact on the effective power. For instance, the primary target audience for this study was novice drivers. Owing to the closure of the university during the 2020 coronavirus pandemic, we did not have access to the same pool of potential participants as we would normally expect, and thus the resultant sample was more heterogeneous. This inevitably introduced additional variance into the data, and may have reduced the potential for obtaining an effect, as a full sample of novice drivers might have been more receptive to this technology-based training, and have a greater skill gap than more-experienced drivers, into which more value could be added.

One further possible reason for the equivocal findings is that the control group may have received some training benefit from the baseline assessment of hazard skill. Prior to the control training condition (the mind-wandering task), these drivers witnessed 20 hazards in short succession in the process of having their baseline performance established (ten hazards in the simulator and ten hazardous precursors in the single-screen prediction test). It is likely that mere exposure to such hazards will lead to driver improvement. If one looks at hazard prediction *improvement* across the pre- and post-training tests, we note that even the control group improves from 62% accuracy to 68%. While this is much less that the VR-trained group who improve from 62% to 75%, it still represents a benefit gained from undertaking the first hazard prediction test (and possibly the simulator drive).

5.4.2 How might training be improved

It is possible that the video-based hazard clips might have been more suitable as training materials. The CGI clips were chosen on the basis that they were simpler and less visually complex than the video-based clips. This provided a scaffolded learning experience, allowing drivers to focus on the key training messages without unnecessary distraction. This is, however, based on the assumption that skills learned in CGI translate to the real world. When looking at those clips that suggested a training benefit, the strongest effects were noted in video clips that almost exactly emulated the core hazard of one of the training clips. Unfortunately, it appears that abstraction beyond these highly similar scenarios was not forthcoming.

This echoes the cognitive literature of visual learning and abstraction. Two extreme viewpoints are: (1) that we learn to predict outcomes on the basis of prototype scenarios from which we abstract rules which we then extrapolate and apply to a range of situations beyond those we have seen (e.g. the Cue Abstraction Model, Juslin et al., 2003); and, alternatively (2) that we predict outcomes based on learned exemplars of situations, so that our understanding of a new event is based on strong similarity to a stored memory trace of a previous encounter (e.g. the Exemplar-Based Model, Nosofsky, 1986). Prototype learning is more flexible, as it allows extrapolation to a wider range of situations (owing to the formulation of rules), though it does require (in order to infer those rules) more than one instance of a prototypical scenario to be witnessed (with two as the minimum; see the Prototype-Based Model, Henriksson, 2019). The exemplar-based models can function with a single instance of a scenario, but the resultant learning is restricted in its application to situations that are very similar, if not identical, to the exemplar.

While we are stretching theoretical models to fit real-world driving dangers (Henriksson's study was based on drawings of fictional bugs/insects rather than realistic driving events), the logic should transfer to our complex on-road scenarios. If we assume this to be the case, we need more instances of hazards of the same general type in terms of the core hazard (e.g. an oncoming car turns across the driver's lane) that nevertheless vary in subtle ways (e.g. lighting, associated speeds, nearby distractions). With the CGI clips, we had only one instance of each hazard (with ten hazards in total). Unfortunately, it would have been too costly to double the number of clips and have even two instances of each hazard. It is also questionable whether multiple versions of the same underlying CGI hazard would afford sufficient variety in regard to subtle cues, to provide the range of prototypes necessary for achieving the best training transfer.

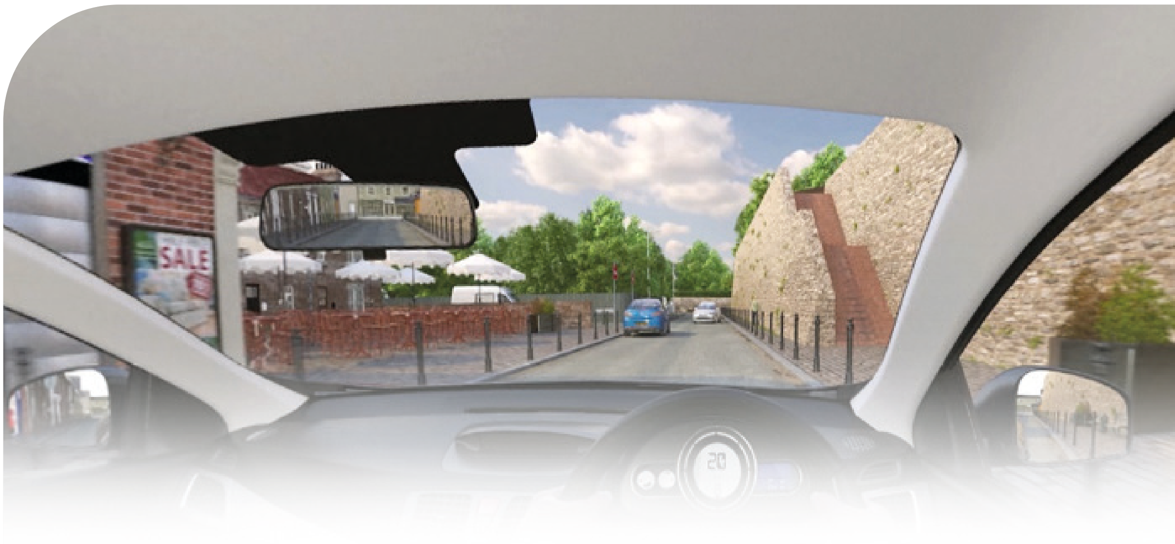
The inevitable suggestion is that more training is required, with repetition of hazards in the training materials. One recent study has trialled the use of repeated exposure to the same hazards and found a training benefit (Kahana-Levy et al., 2019), although as the repetitions were of identical clips, we suggest that they were reinforcing an exemplar-based process, and may have therefore overestimated the potential for training transfer. Instead, we recommend repeating variants of specific hazards (all containing the same core hazard but under a variety of conditions) during training, until drivers can extrapolate underlying rules for judging whether, for instance, an oncoming car is about to turn across their path.

To obtain the required subtlety of cues (and for pragmatic reasons to do with cost), it would make sense to invest further in the generation of video-based clips. This does not mean, however, that we should dismiss CGI clips. We know that they differentiate between safe and less-safe drivers, and they provide a very clear introduction to a particular hazard prototype, in a simplified training environment. A future training program should probably begin with the CGI hazards, but then follow this up with multiple presentations of different variants of the same ten hazards using naturalistic video-based clips. This approach would not be without its difficulties, however: if we are wedded to capturing naturally occurring variants of our CGI hazards, it might require a lot of speculative filming. A more parsimonious approach might be to film many more clips, and then create CGI prototype scenarios based on clusters of video hazards that share an underlying feature.

5.4.3 Conclusions

The results from the training study were not clear-cut but do hold out promise for future iterations of VR training. The training appears to have had a risk-reduction effect on trainees when it comes to their performance on the simulator. Training transfer to the post-training hazard prediction test appears limited, however, to those clips that had the greatest similarity to hazards in the CGI clips, suggesting that our training has tapped into a less-effective exemplar-based mental system. To improve training in the future, we aim to increase the number of similar hazard variants within the training materials, so as to engage a cue abstraction process that will hopefully support any exemplar-based training effects. An ideal solution would start with uncluttered prototypes presented in CGI format before moving onto repetition of subtle variants of the same hazards using more visually complex naturally captured footage.

6. Study 5: A Comparison of Video-based and Computer-Generated Imagery



6.1 Introduction

Studies 1 to 3 have demonstrated that both the CGI and video-based 360-degree tests are able to differentiate between novice and experienced drives, with some evidence to suggest that the 360-degree tests might be more effective at this than the more traditional single-screen methodology. However, in terms of the self-reported data, there is a clear favourite in the 360-degree hazard prediction, at least in terms of participants' ratings of immersion, realism and engagement. Taking this into consideration, the use of a VR-based hazard prediction test for future assessment of drivers shows promise.

Study 4 utilised the CGI clips to create a training intervention designed to improve drivers' hazard awareness when placed in a simulator and in a video-based 360-degree hazard test. There were overall training benefits for drivers' lateral control in the simulator, and there was modest evidence of overall decreases in speed. Training benefits were more sporadic when assessed by the video-based hazard test, with only a select number of clips showing an improvement in responses. These clips had elements that related back, almost identically, to the hazardous scenarios that drivers were exposed to during training. The suggestion arising from this was that more instances of similar hazards should be included in the training package, with increasing levels of complexity (i.e. starting with simplified hazards in CGI, then supporting drivers in translating this learning to variants of the same type of hazard presented in naturalistic video).

As outlined in subsection 4.1, whilst video-based tests dominate the hazard awareness literature, there has been an notable shift in key stakeholders (i.e. government agencies and training companies) from video-based materials to CGI content for hazard awareness assessment. Whilst both presentation modes have theoretical and pragmatic pros and cons when it comes to using them, evidence of whether the end users of the tests prefer CGI or video is relatively sparse within the literature. In a recent study, we presented drivers with an early version of the single-screen CGI test and asked them whether they preferred the CGI to real video footage of a hazard test. The results showed little consensus: some preferred video, some preferred CGI, and some remained ambivalent (Crundall et al., 2021). The final study aimed to address this discord and assess whether participants prefer the use of 360-degree CGI or video for assessment and/or training purposes.

6.1.1 The current study

The ten CGI clips used in Study 3 were matched on their underlying structure to ten of the video-based hazards developed in Study 2. This resulted in two tests, a CGI-based and a video-based test, that all participants completed (counterbalanced across participants). As in Studies 1 to 4, all clips were occluded at hazard onset, with participants being asked “What happens next?” and four text options provided for participants to choose between. Participants viewed both tests within a VR headset. Measures of participants’ ratings of sickness, comfort, realism, immersion and engagement, as well as questions relating to test quality, were compared across the CGI- and video-based 360-degree hazard tests. Although not crucial to the study, hazard prediction accuracy was also analysed across the tests. Whether participants preferred the CGI or video was a non-directional hypothesis.

6.2 Method

6.2.1 Participants

Owing to the very high COVID-19 level (Tier 3) in Nottinghamshire at the time of recruitment, in line with our health and safety protocol we were permitted to recruit only staff and students who were currently working or studying at Nottingham Trent University. Thirty-four participants were recruited (8 learner drivers and 26 experienced drivers). Driving experience was not a factor in the current study, as its primary focus was on driver preferences. Nonetheless, given the likelihood that the skill levels of these two groups differed, we have reported their demographics separately. Two participants (both experienced drivers) were removed from all analyses owing to equipment failure. No participants were removed from this study as a result of sickness. The demographics details of the participants are shown in Table 6.1.

Table 6.1: Demographics of all participants who completed Study 5

Group	N	Gender	Mean age (years)	Mean driving experience (years since passing driving test)
Experienced	24	16 females	32.6	12.3
Learners	8	6 females	32.2	0.0

Source: Authors' own (Study 5)
Note: two participants are not included who were removed due to sickness

6.2.2 Design

The design of this study was a simple comparison of participant preferences across the factor of *hazard medium* (a video-based vs a CGI 360-degree hazard prediction test). Both tests comprised ten clips, providing a total of ten hazardous precursors. The order of tests was counterbalanced across participants and the clips within each test were presented in a random order. Preferences were measured by means of ratings on the CRIE questions, and a series of questions directly referring to test quality, and to suitability both as an assessment test and as the basis of training materials. Additional measures included participants' accuracy on the two tests, and their self-reported sickness measures.

6.2.3 Stimuli

6.2.3.1 The 360-degree video and CGI tests

As we only have ten hazards in the 360-degree CGI test, the same clips used in Study 3 were also used in Study 5. In a departure from Study 3's methodology, rather than presenting the CGI clips sequentially, they were presented in a random order. Though this removed the flow of the CGI route, it was deemed a fairer way to compare this test with the video-based version, as the latter test was composed of isolated, independent clips. The CGI clips were edited such that if two sequential clips were played in 'narrative' order, there was sufficient gap between the end of one clip and the start of another to diminish the feeling of immediate continuity (i.e. the second clip would start much further down the road than the point at which the first clip ended).

The video-based test consisted of ten of the 360-degree video-based hazards developed for Study 2. These were chosen on the basis that the underlying structure of each one was similar to one of the ten CGI clips. As with Studies 2 and 3, all the clips were silent apart from the voice-over providing guidance on where the film car would turn. All clips were occluded at the point of hazard onset, and four options were then presented for participants to choose between.

6.2.3.2 The questionnaires

All participants who completed the study were asked to complete the same demographics questionnaire used in the previous studies prior to attending their laboratory session. Immediately following each of the tests, participants were given the following questions:

A cybersickness question – "*On a scale from 1 to 20, how cybersick do you feel? A rating of 1 reflects no symptoms whatsoever, while a rating of 20 reflects extreme feelings* *of sickness. When you make your judgement, please take into account any feelings of nausea, general discomfort, and stomach problems. Try to ignore other feelings such as nervousness, boredom and fatigue*." A rating of 15 or above was chosen as a threshold for removing participants, though participants were also aware that they had the right to withdraw at any point without explanation. All participants were instructed as to what symptoms to look out for when judging their level of cybersickness.

Clip quality questions – a feedback questionnaire probed a range of participants' opinions of the tests using the same 20-point scale used for assessing cybersickness. Participants were asked to rate their agreement with a series of statements from "strongly disagree" to "strongly agree". The statements focused on the following aspects:

- *Clarity* – "The clarity of the clips was good"
- *Smoothness* – "The clips played smoothly without judder"
- *Complexity* – "The scenes depicted were highly complex (e.g. lots of other road users)"
- *Freedom* – "It was useful to be able to look in all directions during these clips"
- *Assessment* – "These clips would be useful to **test** my hazard skills"
- *Training* – "These clips would be useful to **train** my hazard skills"

CRIE questionnaire: the four CRIE questions were also given, though participants were asked to respond on a 20-point scale to ensure consistency with the other questions.

The cybersickness, quality and CRIE ratings were all presented within the headset, and participants could respond using the VR handheld controller to select a position on the 20-point scale. All questions were presented after both tests. After completing the study, participants were asked two final questions verbally by the researcher regarding which test they thought best reflected their hazard prediction ability:

"Given the current quality of the videos that you have seen, which test do you think offers the best assessment of your ability to predict hazards when driving?" Answers were given on a scale of 1–7, with 1 reflecting "Definitely the Video Test", 4 to mean "Both are equally good", and 7 "Definitely the CGI test".

"If both tests were improved (content, fidelity, etc.), which of the test formats has the potential to be the best in the future?" Answers were given on a scale of 1–7, with 1 reflecting "Definitely the Video Test", 4 "Both are equally good", and 7 "Definitely the CGI test".

6.2.4 Apparatus

Participants viewed both tests on a tetherless Oculus Go VR headset. A bespoke VR app was designed and developed in-house using the Unity development platform. All clips and questions were presented within the headset. Participants selected the correct answer to each clip by pointing the Oculus Go hand controller at their chosen option and pulling the trigger. The same controller was also used to select their ratings on the 20-point sliding scales. The directional voiceovers to accompany the clips were played through the headset speakers. To sterilise the VR headset in between participants a UVC light box (Cleanbox) was used as detailed in Study 4.

6.2.5 Procedure

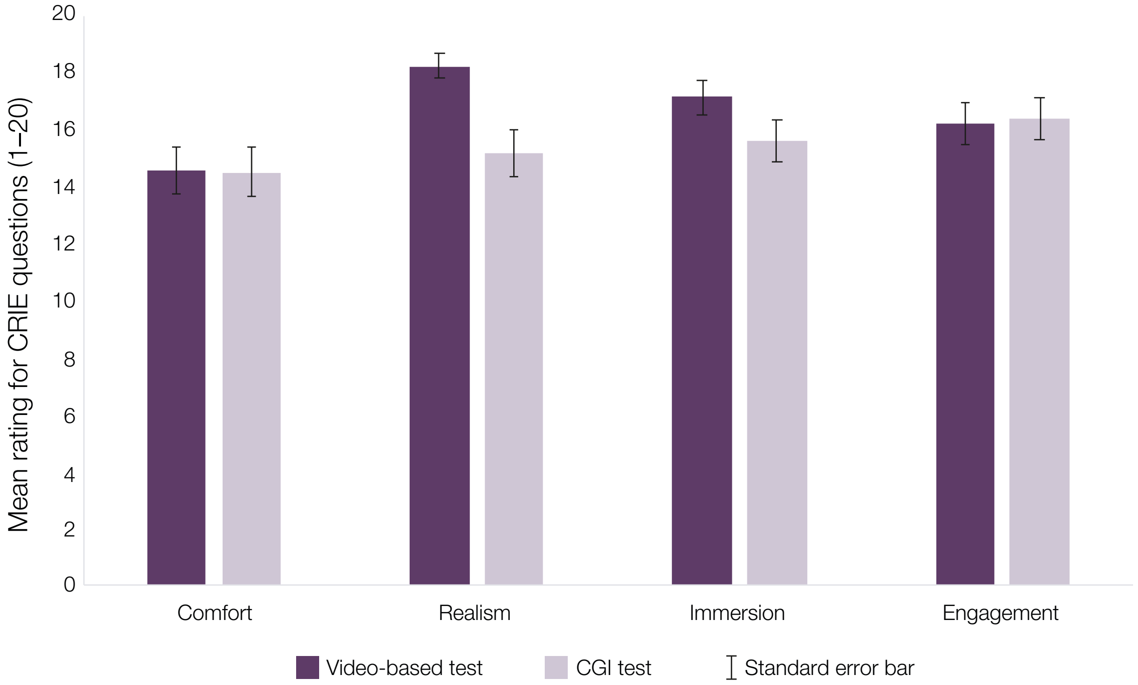
Immediately after signing up to take part, participants were sent a link to a demographics questionnaire via Qualtrics which they were asked to complete prior to their laboratory session. All COVID protocols employed in Study 4 were used in the current study. After signing an online consent form via mobile phone, participants completed both the CGI and video-based tests in a counterbalanced order. Following each test, participants were asked to give their sickness, CRIE and clip quality ratings. These rating questions were displayed in the app on the headsets and participants used the VR controller to select their answers using a sliding scale from 1 to 20. Following completion of both tests, participants were asked two final questions regarding which of the two tests was currently better at assessing their hazard ability, and which test had the possibility of becoming the better of the two with further development. Testing took approximately 30 minutes per participant. All participants received a £10 Amazon voucher for taking part in the study.

6.3 Results

6.3.1 Comfort, realism, immersion and engagement questions

As in Studies 1–3, participants gave ratings for each of the questions regarding CRIE, for both the CGI and video-based hazard prediction tests. Two participants (one experienced driver and one learner) were removed from the CRIE ratings owing to equipment failure. These participants were also removed from the hazard prediction and quality rating questions analyses, but their responses were retained for the final two questions as these responses were captured outside the VR headset. Participants' ratings were entered into a series of paired-samples t-tests. Analysis of the realism question revealed that participants rated the video-based test as significantly more realistic than the CGI test, $t(29) = 3.8$, $p = .001$, (18.3 vs 15.2). Participants' ratings of immersion followed the same pattern, with ostensibly greater immersion for the video-based test, though the difference between the conditions only approached significance, $t(29) = 3.2$, $p = .065$ (17.2 vs 15.6). No other comparisons approached significance (all values of $p > 0.05$). Mean ratings for CRIE questions across the two hazard media can be viewed in Figure 6.1.

Figure 6.1: Average ratings given for each of the four items on the Study 5 CRIE questionnaire

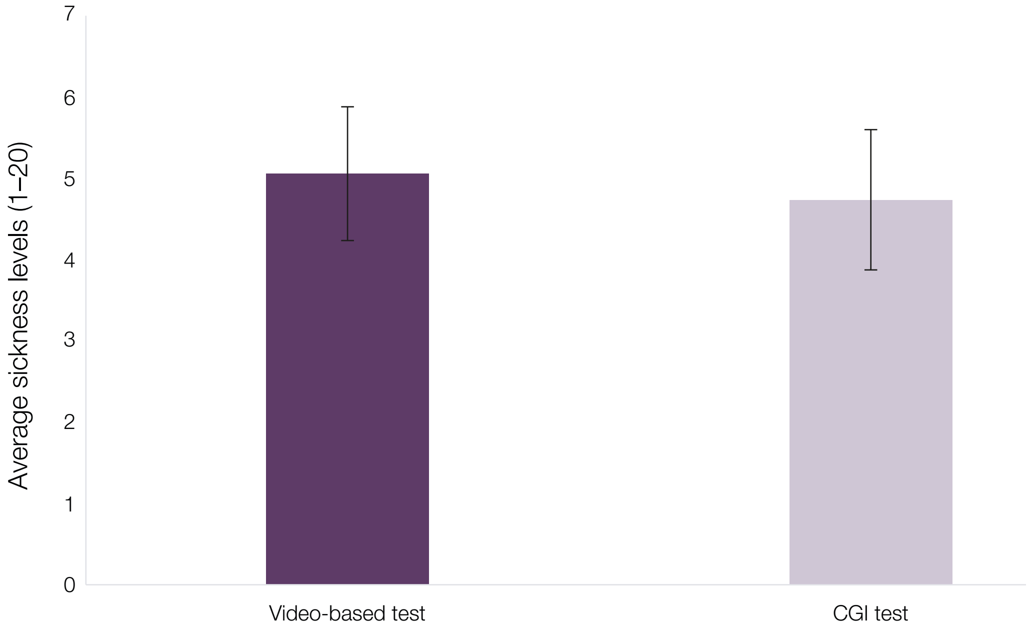


Source: Authors' own (Study 5)
Note: *: p < 0.01

6.3.2 Sickness ratings across the two tests

Participants rated their sickness levels out of a possible 20 (with 20 reflecting extreme sickness) following each test. These ratings were entered into a paired-samples t-test. Analysis of this revealed no significant differences for sickness ratings, *t*(29) = 0.4, *p* = .60 (see Figure 6.2). The sickness rates were consistently low, and no participants were removed from the study.

Figure 6.2: Participants' ratings of sickness for the video and CGI tests of Study 5

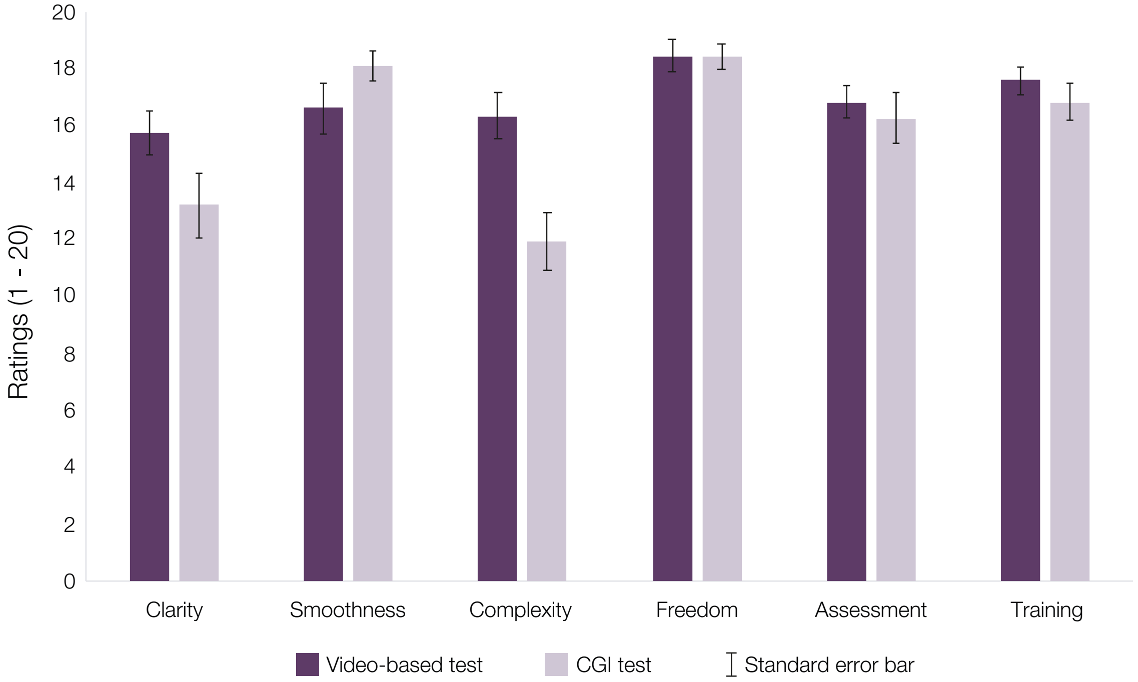


Source: Authors' own (Study 5)

6.3.3 Quality ratings across the two tests

Participants were also asked to provide ratings (1–20) for six questions (the same three for each test) regarding their overall impression of the quality and usefulness of the tests. These ratings were entered into a series of paired-samples t-tests. This revealed that participants rated the video-based test as having greater clarity than the CGI test, $t(29) = 2.3$, $p = .03$ (15.7 vs 13.2), and greater complexity, $t(29) = 5.7$, $p < .001$ (16.3 vs 11.9). No other comparisons approached significance (all values of $p > .05$). The means can be viewed in Figure 6.3.

Figure 6.3: Average ratings given for each of the six questions in Study 5 regarding test quality and use

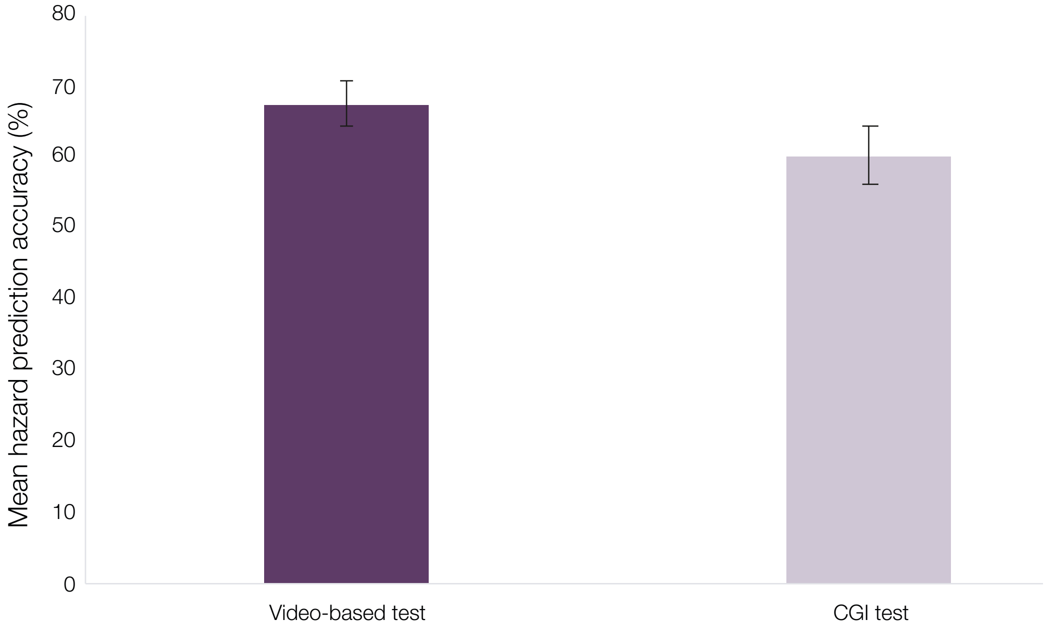


Source: Authors own (Study 5)
Note: * p < .01, ** p < .001

6.3.4 Hazard prediction performance

The percentage of hazards that drivers correctly predicted was compared across the two hazard media using a paired-samples t-test. Although participants appeared to be better at predicting the hazards in the video-based test than the CGI test (67.7% vs 60.3%; see Figure 6.4), this difference did not reach significance (*t*(29) = 1.4, *p* = .17). Given that the video-based test was rated as more complex, it is interesting to note that this did not negatively impact drivers' ability to predict the hazards compared to the less-complex CGI clips.

Figure 6.4: Mean hazard prediction performance for the video and CGI tests of Study 5



Source: Authors' own (Study 5)

6.3.5 Correlation analysis

Correlation analyses were conducted to assess whether accuracy on both the video-based and CGI test were related to the CRIE questions. Correlations were chosen instead of a regression owing to the relatively small number of participants. While 34 participants are adequate for the primary goal of this study (to assess participant preferences), this number is arguably too few to undertake a regression with all our potential predictors. This number was further reduced with the removal of four participants: two due to equipment failure (see subsection 6.3.1), and two who did not provide full CRIE responses.

Separate sets of correlations were calculated for the video-based test and the CGI test (Table 6.2 and Table 6.3). Correlations were undertaken on measures of hazard prediction performance, driving experience, all CRIE questions and all quality questions.

Though comparing drivers on the basis of experience was not an aim of this study (and the sample size reflects that), given that we had a wide range of experience (from learner drivers to highly experienced drivers), it seemed sensible to include experience in the correlation analyses. Driver experience was found to correlate positively with accuracy on the video-based test, but not on the CGI test. This suggests that more-experienced drivers perform better on the video-based hazard prediction, supporting the results noted in Study 2. However, the lack of correlation between experience and accuracy in the CGI does not accord with the findings of Study 3.

Other interesting relationships include those between comfort and smoothness in the video test, and between comfort and clarity in the CGI test. The video clips objectively contain more judder (though participants did not rate this as a significant problem compared to CGI clips), while the clarity of the CGI clips was rated significantly worse than for the video test. It appears that comfort correlates with those factors that a particular test has problems with. The reason for this can be found in the standard deviations of those ratings. Those questions which receive lower ratings, also receive more varied ratings. This provides a spread in the data, which is a prerequisite of a correlation. Where drivers all agree that a clip should score highly on a particular rating, there is little variance in the data to allow for a correlation. Items that correlate with comfort may therefore be indicators of potential problems, especially for the CGI test, where comfort was significantly and negatively correlated with sickness (i.e. participants who report higher sickness also report lower comfort).

The same can be seen with the ratings of realism. Participants rated the video clips as highly realistic, and there was low variance in their responses. The CGI clips, however, were rated as significantly less realistic, though there is lower consensus on this outcome as participant ratings are more varied (i.e. the rating has a higher standard deviation). This variation in response gives rise to the positive correlation with comfort (i.e. those participants who perceive the CGI clips to be less realistic are also likely to rate the test as less comfortable).

The relationship between CGI realism and complexity is worthy of note, suggesting that the significant lack of complexity (subsection 6.3.3) is negatively linked with participants' ratings of realism, and their willingness to see the test used to measure their hazard perception. Indeed, when considering the face validity of the CGI clips as a true assessment of skill, several significant relationships become apparent. This suggests that a CGI test would have to be improved on several dimensions to gain acceptance as an assessment tool by the majority of users. Conversely, only the smoothness of the video clips is significantly linked to the face validity of the clips as an assessment test or in a training context. This could potentially be improved by using a gimbal-mounted camera during filming, or by improvements in post-production editing.

In summary, the correlations tend to favour the video-based test as an *assessment* tool. While some interesting problems remain for the video-based test (why do experienced drivers find the video clips less realistic, and why is immersion positively related to sickness?), there are several points in favour of this test. For instance, in this analysis, the video test was the only one to relate to driving experience. Furthermore, while acceptance of the CGI test is linked to the vagaries of several other factors, the video clips are at the mercy of only their perceived smoothness (which we believe can be improved).

There are, however, fewer problematic relationships between the items and drivers' acceptance of the CGI clips as a *training* tool. Our CGI clips may therefore be more acceptable to participants as a first stage in training (perhaps providing the initial prototype scenario, before further instances are provided in a video medium; subsection 5.4.2).

Table 6.2: Means, standard deviations, and Pearson correlations for measures derived from the video-based test (N = 30)

	Mean	SD	Accuracy	Experience (months)	Sickness	Comfort	Realism	Immersion	Engagement	Clarity	Smoothness	Complexity	Freedom	Assessment
Accuracy	67.7	22.4												
Experience (months)	87.4	92.3	**.410***											
Sickness	5.1	4.5	.229	.200										
Comfort	14.6	4.6	–.007	.092	–.243									
Realism	18.3	2.5	–.276	**–.374***	.359	–.262								
Immersion	17.2	3.4	–.090	–.133	**.441***	–.230	**.455***							
Engagement	16.2	4.3	.200	.160	.037	.311	.088	**.448***						
Clarity	15.7	4.2	–.313	.105	.005	.031	**.512****	.032	–.075					
Smoothness	16.6	4.8	–.315	–.174	–.276	**.396***	.221	–.103	.302	.264				
Complexity	16.3	4.5	–.051	.022	.052	–.356	.050	.270	–.164	.185	.041			
Freedom	18.4	3.2	–.069	–.278	.130	–.156	.204	.210	.006	.090	–.032	.301		
Assessment	16.8	3.0	–.152	–.110	–.195	.083	.347	.306	.352	.338	**.393***	.071	.251	
Training	17.5	2.9	–.259	–.174	–.142	–.052	.247	.330	.243	.291	**.390***	.077	.105	**.816****

Source: Authors’ own (Study 5)

Note: Bold font denotes significant correlations; *: Correlation is significant at the 0.05 level (two-tailed); **: correlation is significant at the 0.01 level (two-tailed).

Table 6.3: Means, standard deviations, and Pearson correlations for measures derived from the CGI test (N = 30)

	Mean	SD	Accuracy	Experience (months)	Sickness	Comfort	Realism	Immersion	Engagement	Clarity	Smoothness	Complexity	Freedom	Assessment
Accuracy	60.3	17.5												
Experience (months)	87.4	92.3	.003											
Sickness	4.7	4.7	.196	–.053										
Comfort	14.5	4.7	–.245	.217	**–.483****									
Realism	15.2	4.6	–.230	.239	.085	**.408***								
Immersion	15.6	4.0	.022	.019	–.083	.169	**.435***							
Engagement	16.4	3.9	.212	.009	–.134	.013	–.024	**.480****						
Clarity	13.2	6.3	–.082	.237	–.124	**.720****	**.602****	.316	.023					
Smoothness	18.1	2.9	.013	.057	–.242	.131	.264	.359	.328	.354				
Complexity	11.9	5.4	**–.362***	.149	–.187	.245	**.542****	.173	.034	.197	.171			
Freedom	18.4	2.4	.164	.001	–.031	–.009	.286	**.532****	**.364***	.058	.231	**.376***		
Assessment	16.2	4.8	–.050	.109	–.118	.283	**.475****	.202	**.416***	**.458***	**.424***	**.518****	.352	
Training	16.8	3.5	.142	–.224	–.186	–.164	–.137	.304	**.461***	–.011	.288	.122	.174	.307

Source: Authors' own (Study 5)

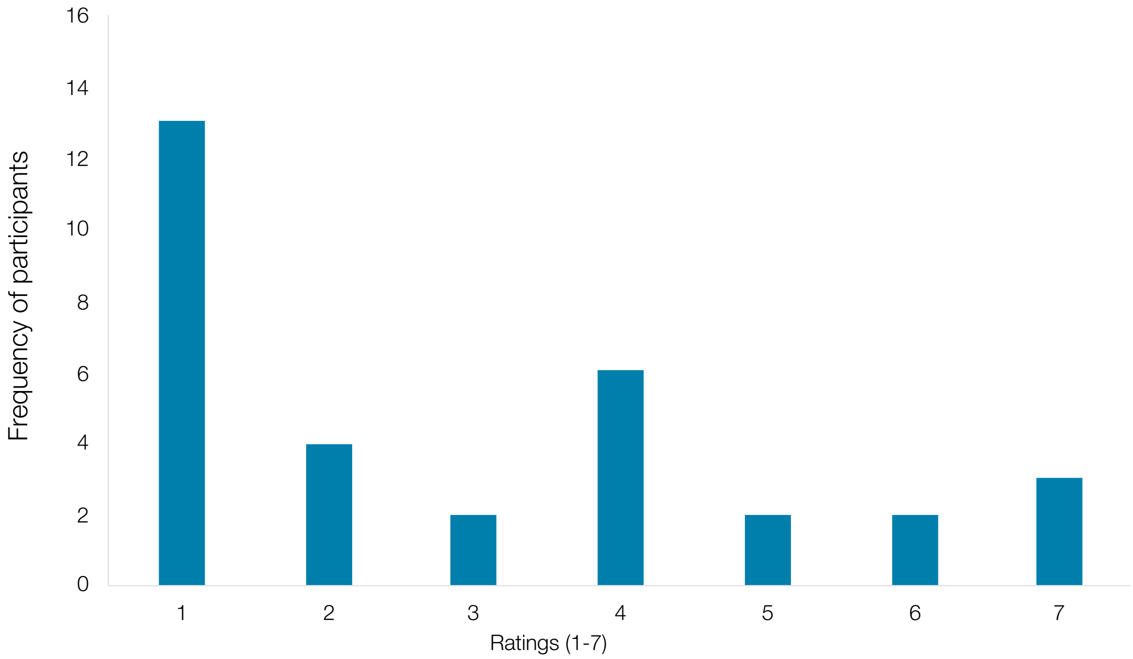
Note: Bold font denotes significant correlations; *: Correlation is significant at the 0.05 level (two-tailed); **: correlation is significant at the 0.01 level (two-tailed).

6.3.6 Participants' evaluation of the tests

Following Study 5, participants were given two evaluation questions regarding which test they thought best reflected their hazard prediction ability. In both questions, participants were asked to give a rating between 1 and 7, with 1 suggesting a strong preference for the video-based test and 7 for the CGI test. The two participants who were excluded from the previous analyses owing to equipment failure were included in this analysis

> ***Question 1: Given the current quality of the videos that you have seen, which test do you think offers the best assessment of your ability to predict hazards when driving?***

Figure 6.5: Frequency of participants' responses when questioned about which test (video or CGI) offers the best assessment of their ability to predict hazards when driving



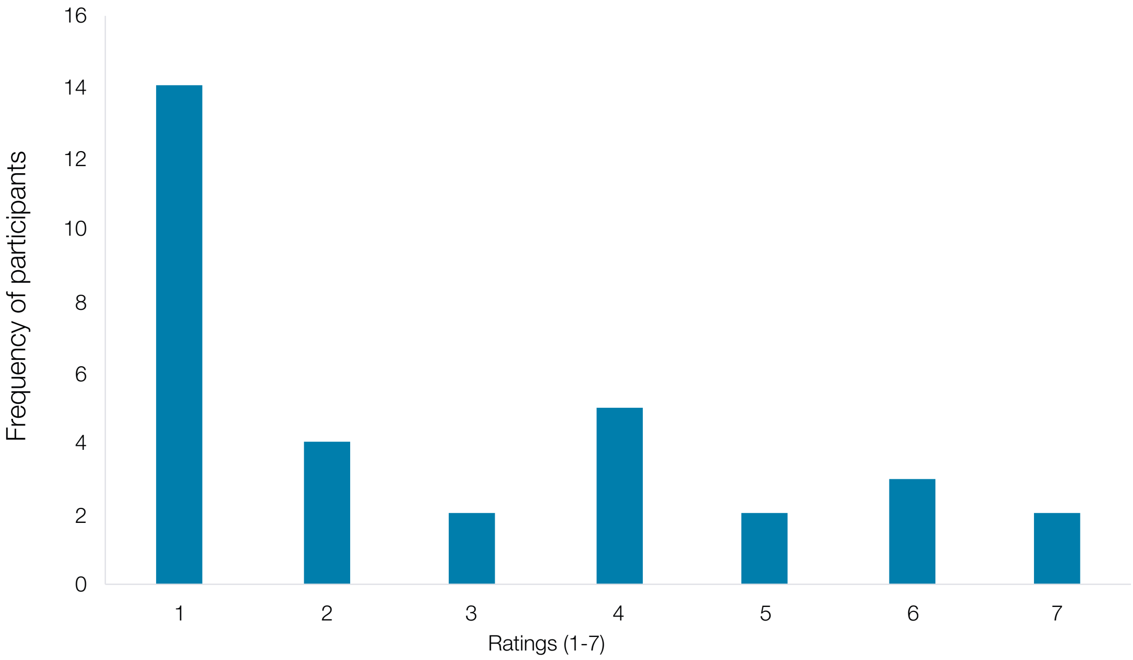
Source: Authors' own (Study 5)

As can be seen from Figure 6.5, 19 (ratings 1 to 3 combined) out of the 32 participants (59%) thought that the video-based test offered the best assessment of their ability to predict hazards compared to seven (ratings 5 to 7 combined) participants (22%) who thought that the CGI test provided the best assessment of their ability to predict hazards. Six participants (19%) thought that both tests were equivalent at assessing their ability to predict hazards.

> ***Question 2: If both tests were improved (content, fidelity, etc.), which of the test formats has the potential to be the best in the future?***

Figure 6.6 shows that 20 (ratings 1 to 3 combined) out of the 32 participants (63%) thought that if the tests were improved then the video-based test would be the best in the future, compared to seven (ratings 5 to 7 combined) participants (22%) who thought that the CGI test had the potential to be the better test if improved. Five participants (16%) thought that both tests would be equivalent if both improved in the future. The pattern across the two tests is extremely similar, with the future possibility of improvement merely moving a small number of participants off the ambivalent fence, onto the side of the video-based test.

Figure 6.6: Frequency of participants' responses when questioned about which test (video or CGI) has the greatest potential to assess hazard prediction if developed further



Source: Authors' own (Study 5)

6.4 Discussion

The primary aim of Study 5 was to identify which test (video or CGI) participants preferred when presented in a 360-degree environment. The data revealed that our drivers rated the video-based test as more realistic, and having greater clarity and complexity than the CGI test. There was no difference in terms of hazard prediction performance between the tests. However, driver experience did correlate significantly with hazard prediction accuracy in the video-based test, but not in the CGI test. Several of these ratings, especially those that were rated lower for one test compared to the other, revealed significant relationships with participants' acceptance of these tests as assessment and training tools. The most problematic number of relationships was noted for the CGI clips' acceptance as a method of assessment. Overall, the video clips won the day in regard to participants' ratings and explicit preferences. However, we believe there is still a role for our CGI clips to play in the initial stages of future training.

It should be noted that any comparison of preferences is based purely on the versions of these tests that we have created. This does not mean that alternative CGI clips, with better clarity and resolution, would not evoke more favourable responses from participants. We already know that single-screen versions of the same CGI clips used in these studies were rated as favourably as video-based clips in a previous study, and that they provided very clear differentiation between groups of drivers on the basis of hazard prediction scores (Crundall et al., 2021). It is highly likely that this success could be replicated with new CGI clips that are designed to overcome some of the participant concerns noted in the current studies.

A particularly interesting finding was that hazard prediction accuracy correlated with driver experience in the video-based test, but not in the CGI test. This suggests that more-experienced drivers perform better on the video test (mirroring the findings of Study 2), yet no such effect was found for the CGI test (ostensibly contradicting the findings of Study 3, and Crundall et al., 2021). We must include a caveat about the lack of correlation of hazard prediction accuracy with experience in the CGI test: only seven novice drivers (after one had been removed owing to equipment failure) contributed to those analyses. Nonetheless, despite low numbers of inexperienced drivers, the relationship did hold for the video clips. It is possible that the video-based test is sensitive to variations of experience even within the group of ‘experienced’ drivers, whereas the CGI test is more responsive to the step change in experience between novices and experienced drivers. Given the potential subtlety of hazardous cues in the video clips, this explanation is highly plausible.

As the video-based test is created from real-world footage, it is not surprising that participants found it to be more realistic, clearer, and more complex – and therefore presumably more akin to hazard prediction on real roads. In corroboration, the evaluation questions regarding which test provided the best assessment of their hazard skills revealed a clear favourite, with over half of the participants (59%) choosing the video-based test, relative to the 22% of participants who thought the CGI test was the best assessment of their hazard skill (with 19% of participants sitting on the fence). In terms of which test participants thought had the best potential for the future (if both content and fidelity were improved), 63% of participants chose the video-based, whereas 22% chose the CGI test (16% of participants rating the tests as equal).

In conclusion, our video-based test was the clear favourite of our participants, though it remains a possibility that future iterations of our CGI clips will improve acceptability. The clarity of the clips was a particular issue, and one that could be easily solved with improved resolution. Study 5 also provided some evidence in favour of our video-based test being more sensitive to driving experience when there are fewer novice drivers in the sample. This concords with the argument that naturally recorded video hazards contain a range of subtle cues that could allow finer distinctions between drivers at the higher end of the experience range.

7. Study 6: Testing Hazard Skills via the Oculus Go Store

7.1 Introduction

In March 2020, Nottingham Trent University closed owing to the coronavirus pandemic and we were not able to collect data for several months. We were unsure when behavioural testing could resume, with Studies 4 and 5 still to run. While exploring alternative methods of data collection, we decided to build an app that we could launch in the Oculus Go Store to collect data for Study 5, reducing the amount of effort required once the university reopened. We began developing the app in May 2020.

In the intervening period, the university partially reopened and permitted us to return to research following a rigorous risk assessment and ethical review. Rapid participant testing allowed us to collect laboratory data for Studies 4 and 5 at speed. This removed the pressure on the VR Oculus app for data collection, though we pressed ahead with the launch regardless. Collecting data through the app was always going to be something of an experiment in itself, as it was unclear whether we could recruit participants in this manner, and how representative of our target audience those recruited might be.

With the resumption of testing in the laboratory coupled with the launch of the app, this provided an excellent opportunity to compare online and laboratory datasets, to assess whether a VR app could generate data similar to that collected under laboratory conditions.

The app was launched as a free download in November 2020. The following sections detail the experience of launching the app, the uptake it received, and the quality of the data that it yielded.

7.1.1 The Oculus Go platform

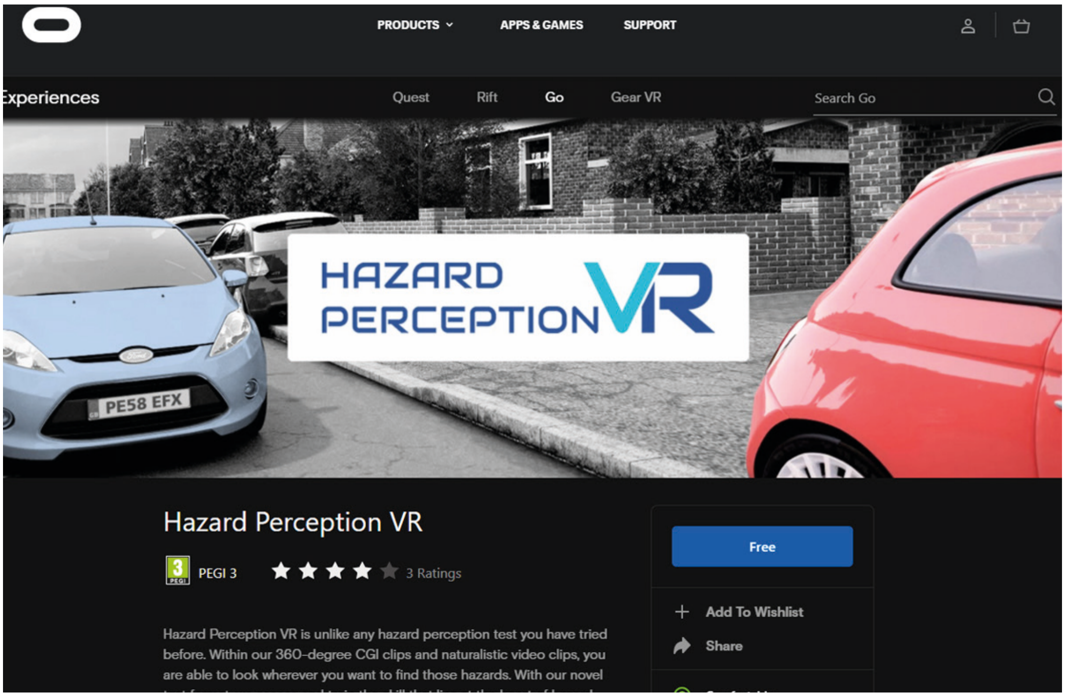
The Oculus Go is an entry-level tetherless VR headset with an LCD screen of 2,560 × 1,440 pixels and a 60 Hz refresh rate. It provides approximately 100 degrees of visual angle, although this depends on the exact positioning on the head. Glasses wearers can use an Oculus Go with an additional fitting. The Go comes with a single control that allows pointing and clicking within the virtual world. The headset is a three-DoF (degrees of freedom) system, which means that it can change what is seen on the screen according to rotational head movements in the three axes (turning the head left or right, looking up or down, and tilting the head from side to side). More expensive headsets use a six-DoF system that tracks the head as it moves in space (as opposed to just rotational movements in three DoF). Thus, with six DoF, one can physically move around a mapped-out area and view objects from different angles. In terms of driving in a virtual car, a three-DoF system will allow the user to look left and right, up and down, and from side to side, but if stopped at a traffic light that is just above the edge of the windscreen, the user cannot lean forward to look up at it. On this basis, six-DoF headsets are more versatile, but the extra DoF become useful only when interacting in a virtual environment that is rendered in real time. With video playback (whether it be CGI or naturalistic video clips) there is no advantage to having a six-DoF system, as the 360-degree content supports only head rotations.

On this basis, we decided to launch on the Oculus Go. Costing £200 at the time, it was the cheapest VR headset that did not rely on inserting a mobile phone into a casing. One downside with choosing the Oculus Go, however, is that it had already been superseded by the six-DoF Oculus Quest. In 2020, the Quest 2 was released and, partway through our development process, Oculus announced that they were withdrawing the Go from sale and closing the Go app store to new apps in December 2020. While the Go app store would continue to be supported until 2022, this was a clear sign that Oculus was moving away from their cheap three-DoF headset, and it was likely that their customer base would follow them with the release of the Quest 2.

For our development process, it was too late to change. Fortunately, we met the December deadline by two weeks, and as of June 2020 our app is still available in the Oculus Go Store.[10] Although its shelf life is limited, it has already achieved several goals such as collecting data from participants, demonstrating market interest through uptake, and allowing us to the be first VR app offering hazard tests to the UK market (see Figure 7.1).

[10] www.oculus.com/experiences/go/3097547357007160

Figure 7.1: The landing page for the Hazard Perception VR app in the Oculus Go app store



Hazard Perception VR

Hazard Perception VR is unlike any hazard perception test you have tried before. Within our 360-degree CGI clips and naturalistic video clips, you are able to look wherever you want to find those hazards. With our novel

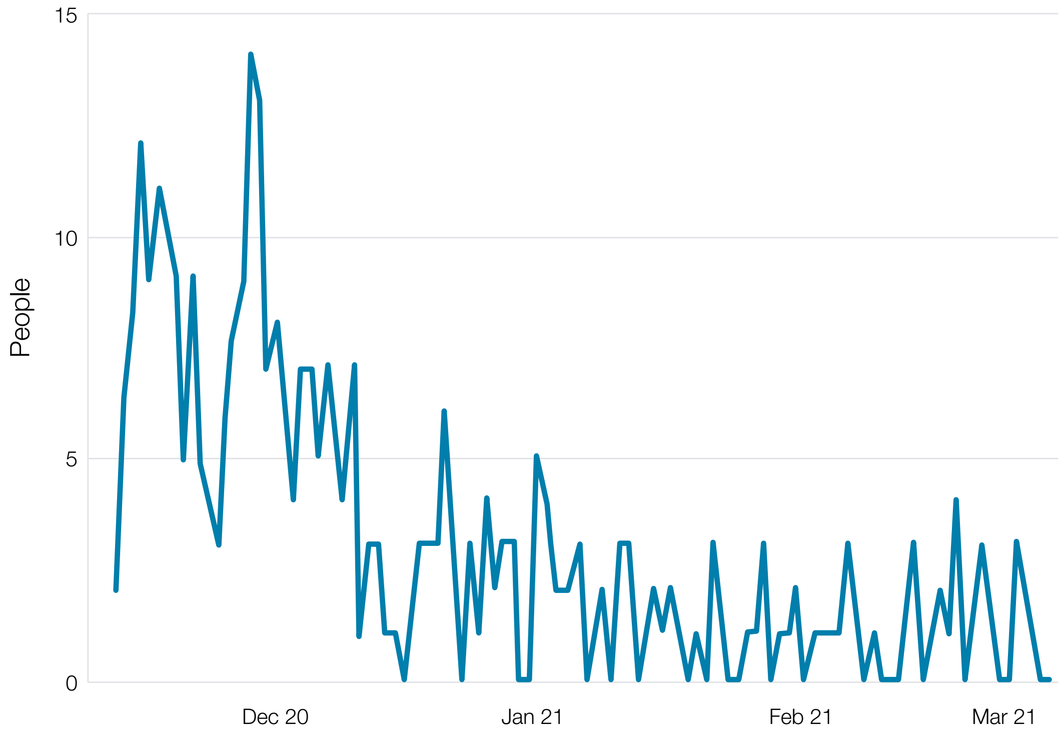
Source: Oculus

7.2 Downloads and uptake

Since its launch in the Oculus Go Store on 12 November 2020, the app has been downloaded and installed 358 times (last counted on 4 March 2021). The largest uptake was in the first two months, but the app continues to be installed on a regular basis over three months later. This total number of installations includes reinstallations and installations on different devices. The total number of unique users is 273 (see Figure 7.2 for the daily number of active users).

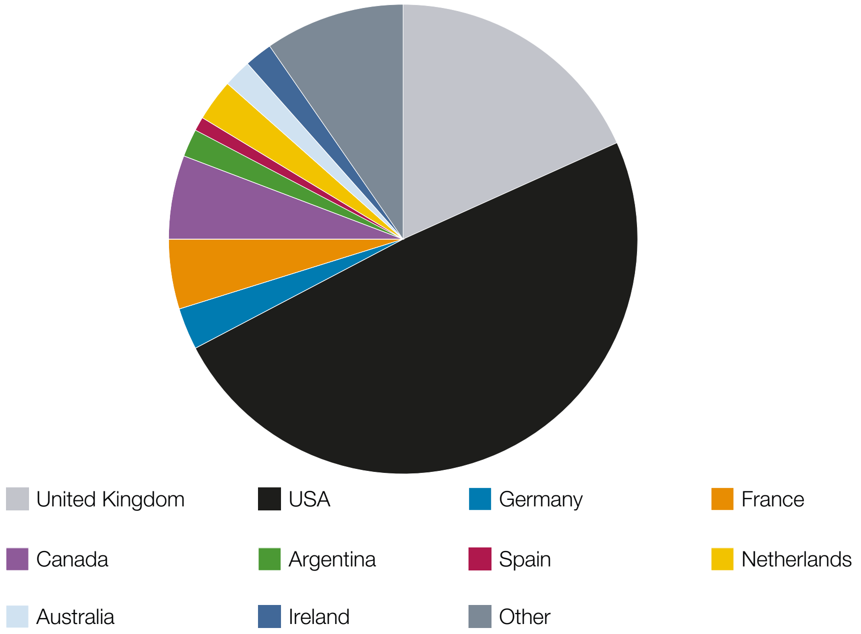
A breakdown of users by country suggests that most of the interest has stemmed from the USA, with 49% of users. The UK has the second highest uptake rate, with 18% of all users (Figure 7.3).

Figure 7.2: The number of active users of the Hazard Perception VR app



Source: Oculus Analytics

Figure 7.3: Countries of origin of active users of the Hazard Perception VR app

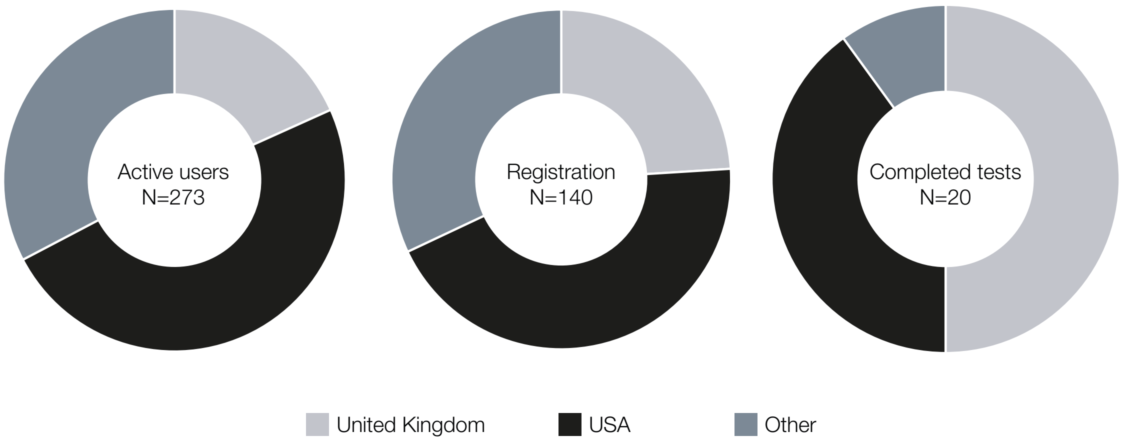


Source: Authors' own (Study 6)

Of the active users, 182 started the registration process. The registration process required users to fill in a web-based registration questionnaire. The app directed users to the questionnaire using a native web browser. Of the 182 users who started the registration process, 140 (85 male, 39 female, and 16 who stated ‘non-binary’ / ‘prefer not to say’) completed the form, giving consent for their data to be used for the study. Eleven users did not give consent, and 31 users initially consented but did not complete the registration form. Twenty of the registered users completed both the CGI test and the video-based test.

The breakdown of the country of origin for active app use, study registration and study completion can be seen in Figure 7.4. Whereas only 18% of the total app users were based in the UK, the percentage of UK residents increased for study registration and study completion to 24% and 50%, respectively. The two participants outside the UK and the USA who completed both tests were from Spain, and Trinidad and Tobago.

Figure 7.4: Breakdown of the country of origin for active use of the Hazard Perception VR app, study registration and study completion



Source: Authors’ own (Study 6)

7.3 Method

The methodology of the study was identical to that of Study 5, with the exception that Study 6 was completely self-administered by the owners of Oculus Go headsets who chose to download the app. Despite the considerable uptake since November, full data was collected from only 13 males, five females and two individuals who responded ‘non-binary’ / ‘prefer not to say’. Their average age was 44.9 years (*SD* = 18.1). Four users reported that they were learning to drive. The average driving experience of the users who had passed their test was 30.3 years (*SD* = 16.5).

All drivers were required to fill in a web-based demographic form (accessed through the browser in the Oculus Go) and had to check a box to provide consent for us to use their data. Access was then granted to both tests used in Study 5 (CGI and video). Users were informed that they had to sit both tests in one sitting for the data to be used in our study. Following each test, they were asked to fill in a cybersickness rating, CRIE ratings and quality ratings.

Once they had completed the tests, a library of feedback clips was unlocked. These clips revealed what the answers were to each clip, and provided some rudimentary voice-over feedback, similar to the expert commentary used in Study 4, but on a smaller scale.

7.4 Results

7.4.1 Comparing computer-generated imagery and video clips within the app

Hazard perception accuracy, CRIE ratings, sickness ratings, and quality ratings produced by app users were compared across the hazard media (video vs CGI) with a series of paired-samples t-tests. None of these comparisons approached significance (all values of $p > .05$). The means can be viewed in Figure 7.5 (see the purple and pink bars). Accepting the caveats of a small and internationally representative sample, it is interesting to note that several differences found in the laboratory data (Study 5) were not apparent in the app data. To investigate this further, we directly compared app data with the laboratory data from Study 5.

7.4.2 A comparison of the lab-based results with app-based results

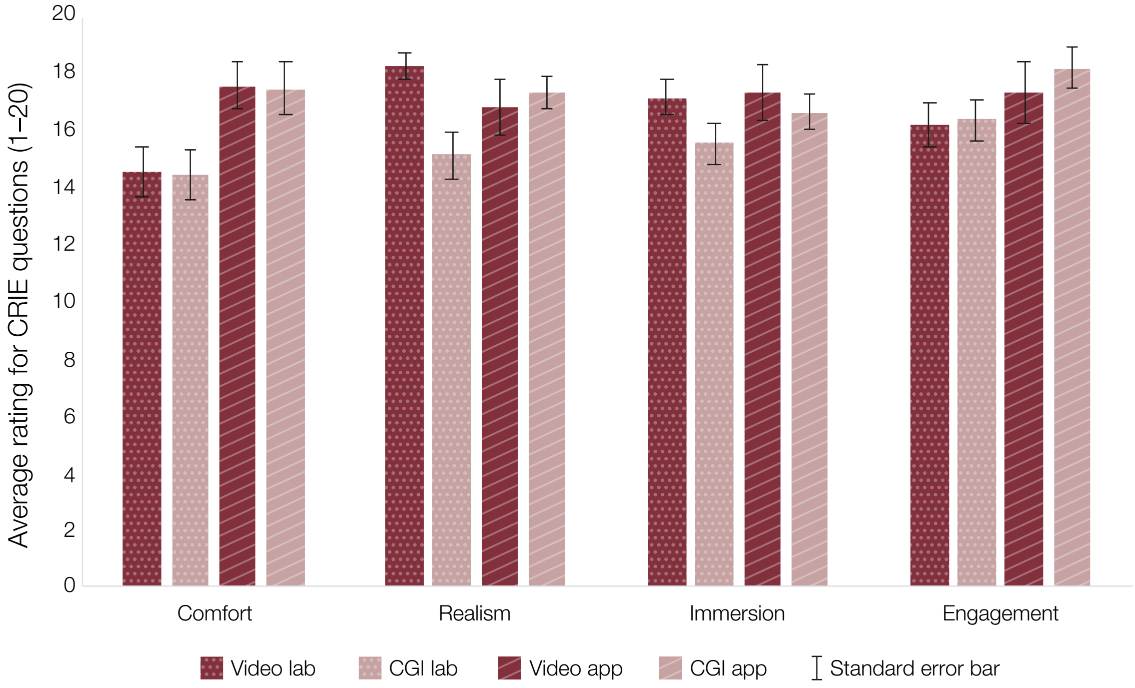
7.4.2.1 CRIE questions

Participants gave ratings for each of the questions regarding CRIE for both the lab and app versions of the tests (see Figure 7.5 for means). Two participants from the app-based version were not included in this analysis owing to missing data. Participants' ratings for each measure were entered into a series of between-subjects 2 × 2 ANOVAs across test delivery method (lab vs app) and hazard media (video vs CGI).

For the comfort question, there was a main effect of test delivery ($F(1, 46) = 7.2$, $MSE = 28.4$, $p = .01$, $\eta^2 = .14$), with the app users rating the tests significantly more comfortable than the participants in the lab (17.6 vs 14.5) regardless of hazard media.

For the realism question, there was an interaction between the factors ($F(1, 46) = 8.8$, $MSE = 8.6$, $p = .005$, $\eta^2 = .16$): although participants in the laboratory rated the videos as more realistic than the CGI clips, this difference disappears in the app data, with mean ratings falling in between the two extremes noted in the laboratory. No other main effects or interactions for immersion and engagement approached significance (all values of $p > .05$).

Figure 7.5: Average ratings given for each of the four items on the CRIE questions for the lab-based tests from Study 5, and the app-based tests from Study 6

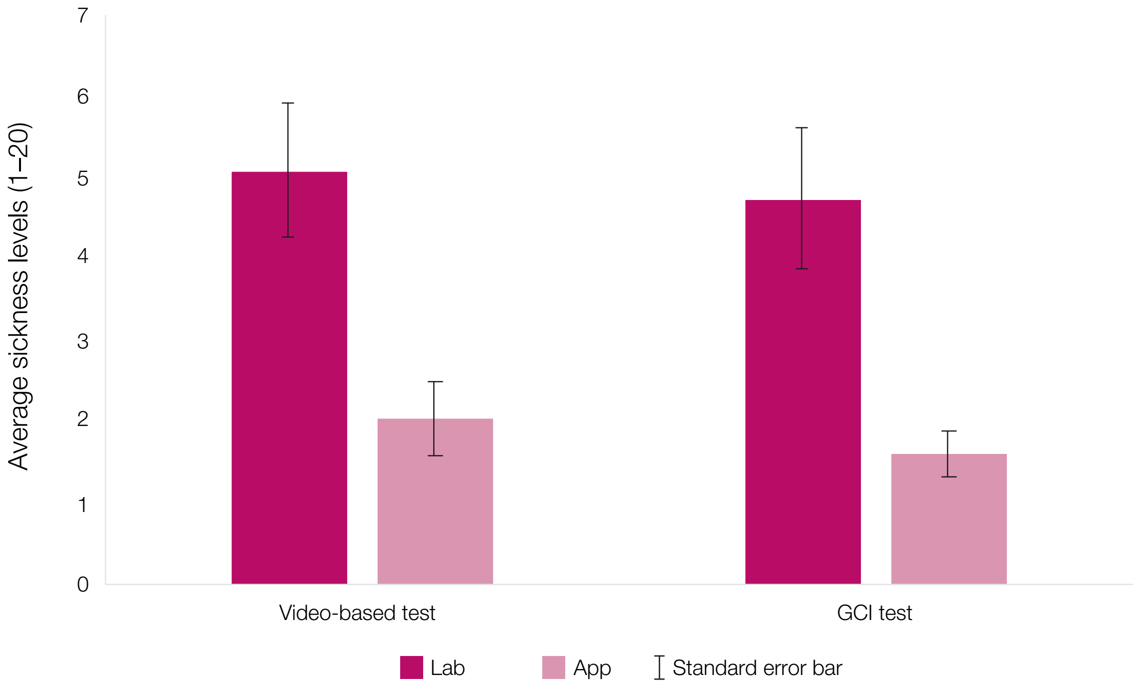


Source: Authors' own (Study 6)
Note: *: p < .05

7.4.2.2 Cybersickness ratings across the lab- and app-based tests

Participants rated their sickness levels on a 20-point scale after undertaking both tests (with 20 reflecting extreme sickness). These ratings were entered into a 2 × 2 ANOVA across test delivery method and hazard media. A main effect of test delivery was noted ($F(1, 48) = 6.9$, $MSE = 23.8$, $p = .01$, $\eta_p^2 = .13$), with the app users reporting significantly lower sickness ratings than the participants in the lab (1.8 vs 4.9) regardless of hazard media. There was no significant difference in sickness ratings between the two tests, nor an interaction between the factors (Figure 7.6).

Figure 7.6: Participants' ratings of sickness for the video and CGI tests of Study 6 across the lab and app



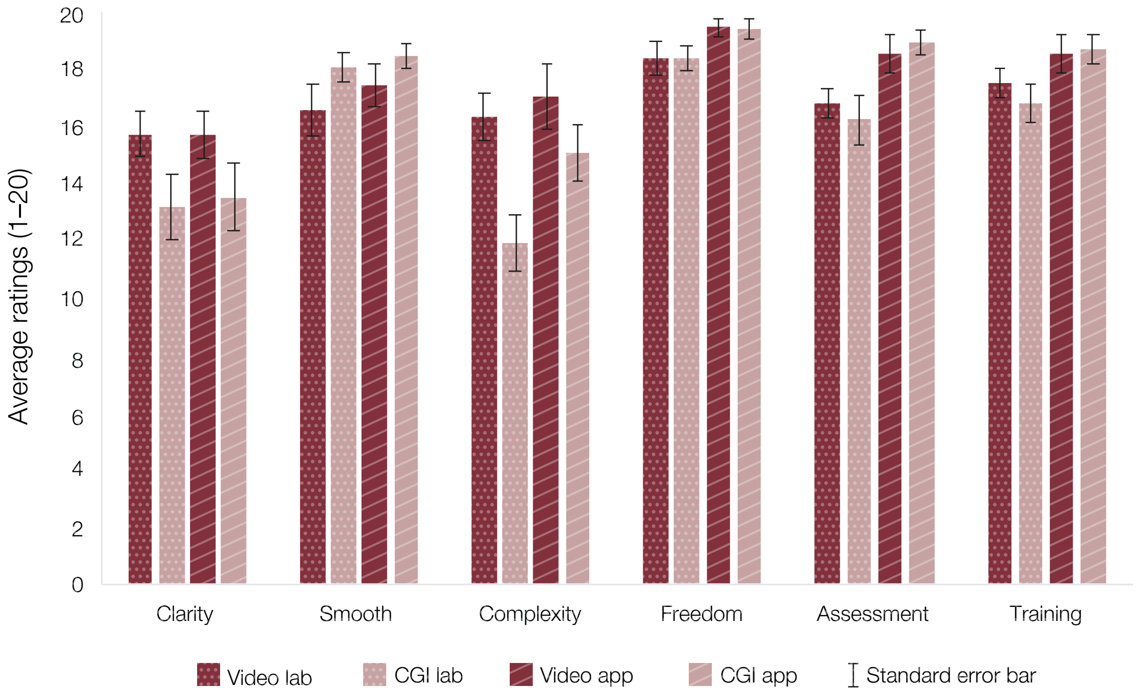
Source: Authors own (Study 6)

7.4.2.3 Participants' quality ratings for the lab- and app-based tests

Participants gave ratings (1–20) for the six questions regarding their overall impression of the quality and usefulness of the CGI and video tests (Figure 7.7). Three participants from the app-based version were not included in this analysis owing to missing data. These ratings were entered into a series of between-subjects 2 × 2 mixed ANOVAs across test delivery method and hazard media.

For the clarity question, there was a main effect of test, $F(1, 45) = 7.4$, $MSE = 16.4$, $p = .009$, $\eta_p^2 = .14$, with participants rating the video clips as having greater clarity than the CGI clips (15.7 vs 13.3) regardless of whether the tests were delivered in the laboratory or in the app. It appears that the preference for the clarity of the video clips that was noted in Study 5 was mirrored in the participants who undertook the study via the app.

Figure 7.7: Average ratings given for each of the six questions in Study 6 regarding test quality and use



Source: Authors' own (Study 6)

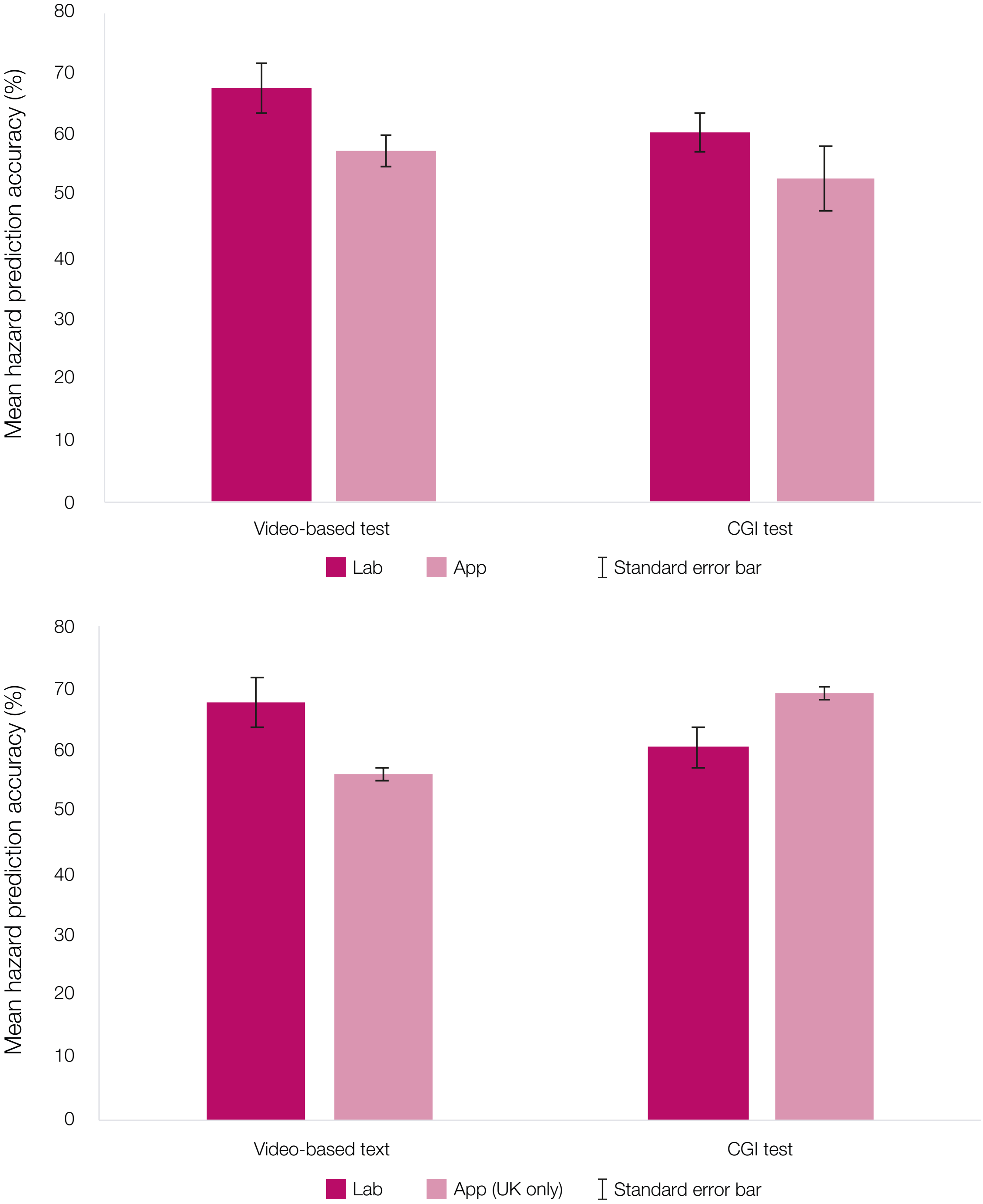
Regarding complexity, there was a main effect of hazard media (*F*(1, 45) = 24.1, *MSE* = 9.2, $p < .001$, $\eta_p^2 = .35$) suggesting that all participants regard the video clips to be more complex than the CGI clips (16.7 vs 13.5). Despite the apparent weakening of this effect in the app data, the interaction did not reach significance (*F*(1, 45) = 3.4, *MSE* = 9.2, $p = .07$, $\eta_p^2 = .07$). Thus, it appears that complexity ratings are relatively consistent between laboratory and app.

For the question regarding how useful participants thought the tests were for *assessing* their hazard skill, there was a main effect of test delivery (*F*(1, 45) = 5.8, *MSE* = 18.6, $p = .02$, $\eta_p^2 = .11$), with app users rating the tests as more useful for assessing their hazard skills than the lab-based users (18.7 vs 16.5), regardless of whether the test was CGI or video. There was a trend for app users to also rate the tests as more useful for *training* their hazard skill than users of the laboratory-based test (18.6 vs 17.2), though this effect failed to reach significance (*F*(1, 45) = 3.5, *MSE* = 13.1, $p = .068$, $\eta_p^2 = .07$).

7.4.2.4 Hazard prediction performance across the lab- and app-based tests

In both the lab and app versions of the tests, participants saw ten clips in both the CGI and video-based tests. Percentage prediction accuracy was compared using a 2 × 2 mixed ANOVA across test delivery method and hazard media. App users scored significantly less than lab-based participants, regardless of hazard media (55.3% vs 64.0%; *F*(1, 48) = 4.7, *MSE* = 395.6, $p = .036$, $\eta_p^2 = .09$). This main effect can be viewed in Figure 7.8 (left panel). The interaction was not significant, and neither was the main effect of hazard media.

Figure 7.8: Hazard prediction performance for the video-based and CGI tests across both test delivery methods (top panel), and with all non-UK app users removed (bottom panel)



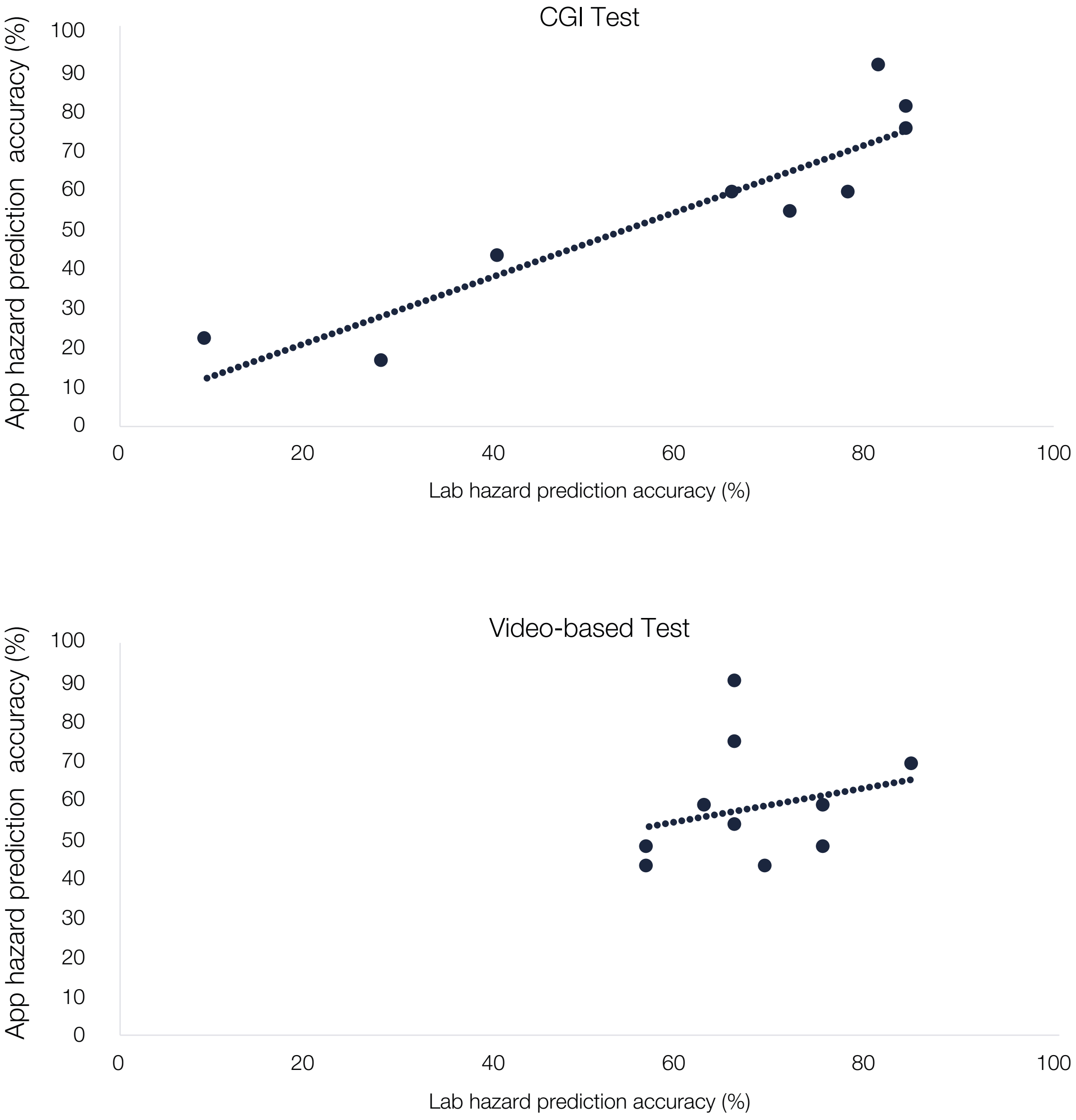
Source: Authors’ own (Study 6)

The app users who contributed data to this analysis included drivers from various countries across the world, with only 50% of participants being UK-based. Research suggests that hazard perception skill can be culturally specific to particular regions, with different social and legal rules influencing both the nature of the hazards and how drivers respond to them (Ventsislavova et al., 2019). Given that both the CGI and video-based test depicted UK roads, it is perhaps not surprising that the lab-based participants outperformed our international sample of app users. To address this potential issue, the analysis was rerun, but using only UK app users (see Figure 7.8, right panel). This analysis revealed an interaction between the factors (*F*(1, 38) = 4.6, *MSE* = 334.2, *p* = .038, η_p^2 = .11). In Study 5, the difference between the performance on the laboratory test between the two hazard media was not significant, despite a slight trend for participants to score better on the video clips. With app users, however, there is a similarly slight effect in the opposite direction, with accuracy favouring the CGI clips. While neither effect might stand on its own, together they produce a crossover interaction. The result suggests that UK app users are perhaps more comfortable with interpreting CGI scenarios than the average drivers who were tested in the laboratory. As these app users are relatively early adopters of VR technology, this greater ease with CGI environments is understandable.

7.4.2.5 Correlations of scores of individual clips across the lab- and app-based tests

To identify whether there was a relationship between performance in the laboratory and performance in the app, accuracy for each clip (the percentage of participants who correctly answer each clip) was correlated across the two test delivery methods. As can be seen from Figure 7.9 (top panel), there was a strong positive correlation between accuracy in the lab-based CGI test and accuracy in the app-based CGI test (*r* = .92, *p* < .001). However, for the video-based test, there was no significant correlation between the lab- and app-based clips (*r* = .30, *p* = .41; see Figure 7.9, bottom panel).

Figure 7.9: Significant correlation between performance on the lab-based and app-based CGI test (top panel), and correlation between performance on the lab-based and app-based video test (bottom panel)



Source: Authors' own (Study 6)

Note: each point on the graphs represents the percentage of participants who responded correctly to that clip

7.5 Discussion

7.5.1 Reception of the app

The app was released on 12 November 2020 with modest social media activity. Our sponsors advertised the app through their networks, and the research team shared the app through social media networks including Twitter. The 350 downloads represent a reasonable level of interest given the low-key advertising, and the announcement that Oculus was discontinuing the Go headset.

There was a considerable drop-off in participants during the progression through the various stages of registration, starting the tests and completing the study. Despite the clear description of the app in the store (i.e. that it was offered as part of a study), it seems clear that many potential users downloaded the app and were then discouraged from further use as a result of the barriers to entry (i.e. filling in a consent form and demographic form before gaining access to the clips).

While the number of users who completed the app-based study was low (we collected fewer datasets than in the laboratory), we presume that many more people who downloaded the app would have become active users if the app was designed primarily for them (rather than for collecting data). Removal of the consent form, demographics form and all the other trappings of an experimental study, would probably create an app with wider active user appeal. If this were then launched on a more modern platform (e.g. the Oculus Quest 2) with a suitable level of advertising, it would be likely to reach a much larger audience.

Of note was the level of interest from the USA. The applicability of the app to the US market might appear questionable given the difference in road rules and street signs between the two countries, and the fact that US cars are driven on the opposite side of the road to the UK. However, many core hazardous elements are likely to translate into the US driving culture, and this may have been of interest to those who initially downloaded it.

The group of participants who completed the whole experiment are perhaps atypical of the standard Oculus user. With an average age of 45, and several decades of driving experience, these drivers would not be considered the natural market for a VR app. We anticipated that learner drivers might feature more than they did, owing to their younger age and likelihood of VR ownership, and because of their need to seek out hazard training in order to pass the UK hazard perception test. Instead, it is probable that these users dropped out at one of the entry barriers, with only four learners completing the study. The majority of the participants who completed the study are therefore likely to have a specific interest in driving-related apps. This should be borne in mind when interpreting the data.

7.5.2 Did the app produce similar data to the laboratory test?

App users responded differently to the tests from the laboratory participants in several regards. First, they reported lower sickness symptoms and greater comfort ratings. As these participants are self-selecting owners of VR headsets, this is not surprising. Repeated use of VR headsets is likely to diminish mild sickness symptoms, while any potential participants who have previously felt severely sick in VR are unlikely to have voluntarily downloaded our app.

App users were also more forgiving of the CGI clips' limited realism that was noted by participants in Study 5. As regular VR users, these participants will be used to engaging in CGI environments, and may be less likely to draw comparisons with the real world than less-seasoned users of the technology. They did, however, report that the clarity and complexity of the CGI clips were lower. These concerns mirror the data received from the laboratory study. Despite this, app users rated both the CGI and video test as being useful as an assessment of their hazard perception skills, more so even than the laboratory participants.

The hazard prediction accuracy scores suggested that app users did not perform as well as laboratory participants, but this was due primarily to the inclusion of international participants who may have had a cultural disadvantage. For those drivers from a non-English speaking country, the selection of the correct multiple-choice option might have had additional associated cognitive demand. Removing the international participants from this sample and comparing the data with Study 5 then revealed an interaction, with UK app users performing slightly better in the CGI test, compared to UK laboratory participants who performed slightly better in the video-based test. This small but significant switch from video to CGI may yet again reflect the greater comfort that app users might have with making decisions in CGI environments.

Correlations between clip performance across the two delivery methods revealed a significant relationship only for the CGI clips. This suggests that clips which tend to receive correct responses in the laboratory will also receive correct responses in the app. Figure 7.9 suggests that the positive relationship in the CGI clips is driven by three clips that are particularly difficult to predict (regardless of whether one sees them in the laboratory or the app).

Overall, it appears that the app users are far more tolerant of the CGI clips than the laboratory users were. This is probably due to the greater amount of time they spend in CGI-based worlds. For them, CGI worlds are likely to take up a greater proportion of their reality than for our laboratory participants. They may therefore be less prone to draw comparisons between a CGI environment and the real world when forming their opinions. There are, however, still two sticking points, as the app users still recognise the limited clarity and complexity that the laboratory participants noted in Study 5. These factors are fixable, however, and these results offer an opportunity to employ CGI clips in future apps. CGI may be particularly useful for multiple international contexts. The strong correlation between laboratory data and app data for CGI clips (which included data from ten international app users) suggests that their simpler structure may translate more easily for international users than the complex video clips, which remain idiosyncratically British.

7.5.3 The future of Hazard Perception VR

The technical lessons learned from the process of designing and publishing an app fall outside the remit of this report, but the experience has been a valuable one. The app itself has been well received, and we believe that if it is repackaged as a training app (rather than an app bogged down by experimental protocols) then it could have considerable take-up from those learner drivers in the UK who have a VR headset.

It is our hope to refresh the app in 2021/22 with a release on the Oculus Quest store. Following lessons learned in the current study, there are clear opportunities to create a training and assessment app that could have a real impact on UK drivers' hazard perception skills. With this in view, we will be seeking further funding to develop more content and to engage with professional app developers to create a more professional product for a wider market.

8. General Discussion

The studies detailed in this report have demonstrated that 360-degree hazard tests are both feasible and worthwhile. Each subsection of this chapter will detail one of the important areas covered in the studies and discuss the ramifications of our findings.

8.1 Cybersickness

We were concerned that our tests would induce high levels of ‘cybersickness’, but this was not the case. Indeed, our favoured test variant (the hazard prediction test) was thought to be particularly likely to evoke cybersickness. The results, however, revealed that the prediction test produced *significantly less-severe symptoms* than the traditional hazard perception test format. Across the five laboratory studies, only 15 participants were removed owing to cybersickness from a total sample of 402 (3.7%). This is a remarkable figure when compared to the literature, where a sickness rate of below 10% is considered a success (Mangalore et al., 2019). Our levels are, however, in line with those of Agrawal et al., (2018). Their study is perhaps the closest in design to our current studies and evoked a similar sickness rate (3%).

It is possible that our low cybersickness rates are attributable to the use of the car interior overlay, which provides a plausibly stable area of the world with which to orient oneself (Prothero, 1998; Prothero & Parker, 2003). Alternatively, the non-interactive nature of hazard tests may reduce cybersickness (in other words, unlike a simulator, the viewer does not provide any steering input). Making a steering movement may prime one to expect a vestibular movement, thus making the mismatch between subsequent visual and vestibular cues more apparent.

Despite the low rates of sickness in our studies, the few serious cases raise problems for the future use of virtual reality (VR) hazard tests in any summative assessment. National tests, for instance, must be designed for inclusive access across the population. If even three or four users per 100 had trouble with such a system, this might pose an access problem for any formal assessments. While an alternative single-screen test could be provided for those people who know they suffer from cybersickness, we have already noted that several participants rejected the notion of cybersickness before succumbing to it. Bearing this in mind, it is difficult to imagine that the ability to opt oneself out would remove all instances of sickness. A further problem lies in the ease of spotting or predicting hazards in a VR headset. We have reported that accuracy tends to be higher in the VR headset (at least with video-based clips), most probably owing to the increased size of the visual scene. To remove cybersick drivers from a formal VR test and provide them with an identical – but *harder* – single-screen test, might be viewed as discriminatory. While there are ways to better equate VR and single-screen tests (e.g. using a large curved single screen), until future VR hardware and software can almost eliminate cybersickness, any assessments are best kept as formative rather than summative.

8.2 Test efficacy

One method of demonstrating the validity of a hazard test is to show that it can differentiate between driver groups who are likely to differ in terms of their on-road risk. We chose driving experience as our measure of risk, as it is well documented that inexperienced drivers (especially those within 12 months of passing their test) constitute a high-risk category. Collision statistics typically show the crash risk of drivers who have only recently passed their test to be at least three times as great as that of the average driver (e.g. Underwood, 2007). Other studies have used self-reported crash history, though these measures are susceptible to self-reporting biases, and can be very few and far between, even for the very worst drivers (e.g. Crundall & Kroll, 2019, Horswill et al., 2020).

We were concerned that the 360-degree tests might reduce the performance gap between novice and experienced drivers. The larger image size of hazards in the VR headset might, for instance, have improved all participants' scores to the extent of creating a ceiling effect that no longer differentiates between the groups. Alternatively, the 360-degree clips might have improved differentiation between driver groups, by furnishing the worst drivers with more options to be wrong (because there are many more wrong places to look in a 360-degree clip, and the cost of looking in the wrong place is greater, as the eccentricity between an incorrect fixation and the hazard precursor is likely to be larger in the VR headset).

The results demonstrated that both our 360-degree tests and the single-screen tests combined to give a significant result, with experienced drivers outperforming the novice drivers. However, neither performance on the video tests, nor the computer-generated imagery (CGI) tests, produced a significant interaction. Such interactions are necessary to conclusively report whether the 360-degree tests were better than the single-screen tests at separating our driver groups. Despite the lack of interaction, both Study 2 and Study 3 revealed the same pattern of group means, with an apparent improvement in the differentiation of driver groups in the 360-degree tests. Pre-planned comparisons confirmed that the difference between novice and experienced drivers was driven by the 360-degree clips in both studies. While we cannot claim strong evidence to crown the 360-degree tests as the winner, the circumstantial evidence is highly suggestive of this fact. At the very least we can say we have no evidence that the VR-based tests were worse than the single-screen tests, and there is slight evidence to suggest that they might be better.

It is highly likely that future iterations of these tests will result in yet greater superiority of the VR method. None of our video clips actually contained a hazard that originated outside of the cone of vision that one would have on a single-screen test. In previous studies, we have recorded clips that contained natural hazards which would have benefited from a 360-degree view, such as undertaking and overtaking hazards (Ventsislavova et al., 2019), and approaching motorcycles at a T-junction (Crundall et al., 2012b). If similar clips had been captured during our initial recording of footage, then the interactions may have tipped into significance, turning weak evidence into strong evidence. Unfortunately, it is difficult to predict what natural hazards might be caught when filming, and we would have to commit more resources to the filming task to ensure that a wider range of hazards are caught. With CGI it would be easier to design hazards that could benefit from a 360-degree view, though these come with the caveat of potential design biases.

8.3 Participants’ views

The overwhelming response from participants was positive towards the VR-based tests. They rated both the video and CGI tests as more realistic, immersive and engaging when presented in this format. That the 360-degree tests would beat the single-screen tests on realism and immersion seems obvious; however, it was always possible that these measures could have decreased. For instance, the larger image in the VR headset meant that the pixels per inch count was reduced compared to the single-screen clips. This could have reduced feelings of realism. Furthermore, as the 360-degree clips aim to copy the real world, this may have induced a more stringent comparison between the test and reality. Fortunately, this proved not to be the case.

Even if VR superiority in ratings of realism and immersion were forgone conclusions, this does not automatically equate to higher ratings of engagement. Nonetheless, our drivers reported higher engagement when in the VR headset. Furthermore, when asked if the 360-degree tests were suitable for assessment and training of hazard awareness, mean ratings varied between 16.2 and 17.5 on a 20-point scale. These ratings show a strong belief in the new tests among participants.

8.4 Computer-generated imagery or video?

The distinction between CGI and video is confounded by the planned vs naturalistic nature of the hazards depicted. Unfortunately, we cannot disentangle presentation mode from the designed/natural hazard debate. Comments below should be viewed in this light.

Participants explicitly favoured the video clips (Study 5). The CGI clips were reported in Study 5 to have less clarity, which may have contributed to the lower comfort ratings reported in Study 3. Clarity can, however, be improved with increased resolution, though this will increase development costs. It will also increase file sizes. While this is not a problem for a laboratory study using a high-specification system, it is more of an issue when creating an app for distribution through the Oculus Go Store. File size limits must be overcome with the assistance of professional developers before higher-resolution CGI clips can be implemented in the app.

The CGI clips were also rated as having lower complexity. When creating a CGI world, every distracting element has to be thought of, designed and programmed. This inevitably leads to sparser imagery than is provided by video clips. Unfortunately, more realistic levels of complexity may be vitally important in capturing the realism of the scene and providing a training environment that will allow transfer to the real world.

The video clips have their own problems, however. We were concerned that the judder on the footage (even after minimisation via image stabilisation software and post-production editing) would have a negative effect on comfort, realism, immersion and engagement. Participants did not, however, seem overly concerned. While their comfort ratings were lower on the video-based VR test than the single-screen version (Study 2), this difference was not significant. In Study 5, there was a slight trend for participants to report lower smoothness ratings for the video than to the CGI, but again this was not significant.

Despite participant preferences, both the video and CGI tests successfully differentiated between our driver groups when used as assessment tests. Furthermore, stakeholders can rest easy in their pragmatic decisions to choose CGI over video for assessment purposes based on this data. For training purposes, however, the findings are less persuasive.

8.5 Training

The training study suggested that our hazard training reduced drivers' willingness to take risks. They adopted slower speeds on the post-training simulated route. Perhaps because of this, they also reduced their lateral variability. The VR-trained group also reduced their steering wheel error, suggesting that they were less likely to swerve or weave in their lane.

Any positive impact of training on the subsequent 360-degree hazard prediction test was limited to a handful of clips that showed clear overlap between hazard content in the assessment and training clips. This suggests that only 'near transfer' of training was supported.

Two potential reasons for this were suggested. First, the low complexity of the CGI clips may have limited the transfer of learning to more-complex situations.

The CGI clips were chosen as the basis of the training intervention because this lack of complexity scaffolded the learning process, allowing drivers to focus on the teachable moments. It is possible, however, that the simplicity of these clips did not prepare drivers for transferring and extrapolating this knowledge to other hazardous scenarios in much more visually rich environments.

A second possibility is that showing a single instance of a particular hazard will support only exemplar-based learning. To encourage the richer process of cue abstraction from prototypes, we should ideally train drivers on multiple instances of the same type of hazard, all with slight variations in context, to ensure that our trainees extract sufficient guidance to apply the prototype cues to a range of other potentially hazardous situations.

A video/CGI hybrid training package was proposed, where the less-complex CGI clips are used to initially introduce trainees to a particular type of hazard, and are followed by video-based variants of that hazard to support their transfer of knowledge from a simplistic exemplar to a range of naturalistic events.

8.6 Conclusions

In conclusion, these studies have demonstrated that it is feasible to present a 360-degree hazard prediction test in VR, while limiting cybersickness to a what amounts to a small minority of people when compared to published studies in the field. A 360-degree test can be as effective as single-screen test in differentiating safe and less-safe driver groups, and we have evidence to suggest that under certain conditions such tests might be more effective. With further iteration, the superiority of the VR presentation mode is likely to become ever more apparent. While even occasional instances of cybersick participants might be enough to prevent a VR test being used at a national level, such tests could be invaluable at identifying the training needs of drivers.

The benefits of using VR to train hazard awareness are harder to demonstrate, however. Improvements on subsequent hazard prediction performance appear to be limited to those assessment scenarios that are very similar to the training scenarios. We have recommended an iteration to future training efforts that will build on the evidence here, and hopefully improve future training benefits.

Finally, one of the strongest effects to come from these studies is that participants have clear preferences for the 360-degree tests over single-screen versions. Their enthusiasm for VR assessment and training, in terms of perceived realism, immersion and engagement, is arguably reason enough to pursue this route to improved driver safety. While we cannot separate participants' enthusiasm springing from the perceived benefits of VR from that caused by the novelty of the presentation mode, such increased levels of initial engagement may even encourage some drivers, who might not have previously considered it, to undertake voluntary training in the privacy of their own VR headset.

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(33): 20–30.

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)*. In A. Kemeny (ed.), Proceedings of the *DSC 2014 Europe Driving Simulation Conference*. Paris: Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux. pp. 65–71.

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.

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(3): 283–292.

Bos, J. E., Bles, W. & Groen, E. L. (2008). *A Theory on Visually Induced Motion Sickness*. Displays, 29(2): 47–57.

Brooks, J. O., Goodenough, R. R., Crisler, M. C., Klein, N. D., Alley, R. L., Koon, B. L., Logan, W. C., Ogle, J. H., Tyrrell, R. A. & Wills, R. F. (2010). *Simulator Sickness During Driving Simulation Studies*. Accident Analysis & Prevention, 42(3): 788–796.

Bruce, C. R., Unsworth, C. A., Dillon, M. P., Tay, R., Falkmer, T., Bird, P. & Carey, L. M. (2017). *Hazard Perception Skills of Young Drivers with Attention Deficit Hyperactivity Disorder (ADHD) Can Be Improved with Computer Based Driver Training: An exploratory randomised controlled trial*. Accident Analysis & Prevention, 109: 70–77.

Cassavaugh, N. D., Domeyer, J. E. & Backs, R. W. (2011). *Lessons Learned Regarding Simulator Sickness in Older Adult Drivers.* Lecture Notes in Computer Science*,* 6767: 263–269.

Castro, C., Padilla, J. L., Roca, J., Benítez, I., García-Fernández, P., Estévez, B., López-Ramón, M. F. & Crundall, D. (2014). *Development and Validation of the Spanish Hazard Perception Test*. Traffic Injury Prevention, 15(8): 817–826.

Castro, C., Ventsislavova, P., Peña-Suarez, E., Gugliotta, A., Garcia-Fernandez, P., Eisman, E. & Crundall, D. (2016). *Proactive Listening to a Training Commentary Improves Hazard Prediction*. Safety Science, 82: 144–154.

Chapman, P. R. & Underwood, G. (1998). *Visual Search of Driving Situations: Danger and experience*. Perception, 27(8): 951–964.

Chapman, P., Underwood, G. & Roberts, K. (2002). *Visual Search Patterns in Trained and Untrained Novice Drivers*. Transportation Research Part F: Traffic Psychology and Behaviour, 5F(2): 157–167.

Classen, S., Bewernitz, M. & Shechtman, O. (2011). *Driving Simulator Sickness: An evidence-based review of the literature*. American Journal of Occupational Therapy, 65(2): 179–188.

Crundall, D. (2016). *Hazard Prediction Discriminates Between Novice and Experienced Drivers*. Accident Analysis & Prevention, 86: 47–58.

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., Kroll, V., Goodge, T., and Griffiths, M., (2019). *Assessing the Potential of Mindfulness Training in Improving Driver Safety: Final Report for the Road Safety Trust.* Retrieved June, 2021, from https://www.roadsafetytrust.org.uk/reports-and-publications.

Crundall, D., Andrews, B., Van Loon, E. & Chapman, P. (2010). *Commentary Training Improves Responsiveness to Hazards in a Driving Simulator*. Accident Analysis & Prevention, 42(6): 2117–2124.

Crundall, D., Chapman, P., Phelps, N. & Underwood, G. (2003). *Eye Movements and Hazard Perception in Police Pursuit and Emergency Response Driving*. Journal of Experimental Psychology: Applied, 9(3): 163–174.

Crundall, D., Chapman, P., Trawley, S., Collins, L., van Loon, E., Andrews, B. & Underwood, G. (2012a). *Some Hazards Are More Attractive Than Others: Drivers of varying experience respond differently to different types of hazard*. Accident Analysis & Prevention, 45: 600–609.

Crundall, D., Crundall, E., Clarke, D. & Shahar, A. (2012b). *Why Do Car Drivers Fail to Give Way to Motorcycles at T-Junctions?* Accident Analysis & Prevention, 44(1)*:* 88–96.

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.

Crundall, D. E. & Underwood, G. (1998). *Effects of Experience and Processing Demands on Visual Information Acquisition in Drivers*. Ergonomics, 41(4): 448–458.

Dogan, E., Steg, L., Delhomme, P. & Rothengatter, T. (2012). *The Effects of Non-Evaluative Feedback on Drivers' Self-Evaluation and Performance*. Accident Analysis & Prevention, 45: 522–528.

Duh, H. B., Parker, D. E. & Furness, T. A. (2001). *An "Independent Visual Background" Reduced Balance Disturbance Evoked by Visual Scene Motion: Implication for alleviating simulator sickness*. Paper presented at the *SIG CHI 2001 Conference on Human Factors in Computing Systems*, Seattle, WA, USA, March 31–April 5 2001. pp. 85–89.

Duh, H. B., Parker, D. E. & Furness, T. A. (2004a). *An Independent Visual Background Reduced Simulator Sickness in a Driving Simulator*. Presence: Teleoperators and Virtual Environments, 13(5): 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(1): 142–153.

Fisher, D. L., Laurie, N. E., Glaser, R., Connerney, K., Pollatsek, A., Duffy, S. A. & Brock, J. (2002). *Use of a Fixed-Base Driving Simulator to Evaluate the Effects of Experience and PC-Based Risk Awareness Training on Drivers' Decisions*. Human Factors, 44(2): 287–302.

Fisher, D. L., Narayanaan, V., Pradhan, A. K. & Pollatsek, A. (2004). *Using Eye Movements in Driving Simulators to Evaluate Effects of PC-Based Risk Awareness Training.* Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48(19): 2266–2270.

Fisher, D. L., Pollatsek, A. P. & Pradhan, A. (2006). *Can Novice Drivers Be Trained to Scan for Information that Will Reduce Their Likelihood of a Crash?* Injury Prevention, 12(Suppl. 1): i25–i29.

Fisher, D. L., Pradhan, A. K., Pollatsek, A. & Knodler, M. A. (2007). *Empirical Evaluation of Hazard Anticipation Behaviors in the Field and on Driving Simulator Using Eye Tracker*. Transportation Research Record, 2018(1): 80–86.

Forster, Y., Paradies, S., Bee, N., Bülthoff, H., Kemeny, A. & Pretto, P. (2015). *The Third Dimension: Stereoscopic displaying in a fully immersive driving simulator*. Paper presented at the *DSC 2015 Europe Driving Simulation Conference,* Tübingen, Germany, 16–18 September 2015. pp. 25–32.

Fuller, R. (2011). Chapter 2–Driver control theory: From task difficulty homeostasis to risk allostasis. In B. E. Porter (ed.), *Handbook of Traffic Psychology* (pp. 13–26). San Diego, CA: Academic Press.

Golding, J. F. (2006). *Motion Sickness Susceptibility*. Autonomic Neuroscience, 129(1–2): 67–76.

Henriksson, M. P. (2019). *Cue Abstraction and Ideal Prototype Abstraction in Estimation Tasks*. Journal of Cognitive Psychology, 31(1): 76–91.

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.

Horswill, M. S., Kemala, C. N., Wetton, M., Scialfa, C. T. & Pachana, N. A. (2010). *Improving Older Drivers' Hazard Perception Ability*. Psychology and Aging, 25(2): 464–469.

Horswill, M. S., Taylor, K., Newnam, S., Wetton, M. & Hill, A. (2013). *Even Highly Experienced Drivers Benefit from a Brief Hazard Perception Training Intervention*. Accident Analysis & Prevention, 52, 100–110.

Isler, R. B., Starkey, N. J. & Williamson, A. R. (2009). *Video-Based Road Commentary Training Improves Hazard Perception of Young Drivers in a Dual Task*. Accident Analysis & Prevention, 41(3): 445–452.

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.

Juslin, P., Jones, S., Olsson, H. & Winman, A. (2003). *Cue Abstraction and Exemplar Memory in Categorization*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(5): 924–941.

Kahana-Levy, N., Shavitzky-Golkin, S., Borowsky, A. & Vakil, E. (2019). *The Effects of Repetitive Presentation of Specific Hazards on Eye Movements in Hazard Perception Training, of Experienced and Young-Inexperienced Drivers*. Accident Analysis & Prevention, 122: 255–267.

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. Haas (eds), *Virtual and Adaptive Environments: Applications, Implications, and Human Performance Issues* (pp. 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, 41(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*. International Journal of Aviation Psychology, 3(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, S., Lee, S., Kala, N., Lee, J. & Choe, W. (2018). *An Effective FoV Restriction Approach to Mitigate VR Sickness on Mobile Devices*. Journal of the Society for Information Display, 26(6): 376–384.

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.

Krishnan, A., Samuel, S., Yamani, Y., Romoser, M. R. E. & Fisher, D. L. (2019). *Effectiveness of a Strategic Hazard Anticipation Training Intervention in High Risk Scenarios*. Transportation Research Part F: Traffic Psychology and Behaviour, 67: 43–56.

Langham, M., Hole, G., Edwards, J. & O'Neil, C. (2002). *An Analysis of 'Looked but Failed to See' Accidents Involving Parked Police Vehicles*. Ergonomics, 45(3): 167–185.

Lappi, O., Rinkkala, P. & Pekkanen, J. (2017). *Systematic Observation of an Expert Driver's Gaze Strategy: An on-road case study*. Frontiers in Psychology, 8: 620.

Lee, S. E., Olsen, E. C. B. & Simons-Morton, B. (2006). *Eyeglance Behavior of Novice Teen and Experienced Adult Drivers*. Transportation Research Record, 1980(1): 57–64.

Lim, P. C., Sheppard, E. & Crundall, D. (2014). *A Predictive Hazard Perception Paradigm Differentiates Driving Experience Cross-Culturally*. Transportation Research Part F: Traffic Psychology and Behaviour, 26: 210–217.

MacQuarrie, A. & Steed, A. (2017). *Cinematic Virtual Reality: Evaluating the effect of display type on the viewing experience for panoramic video*. In Suma, E. et al. (eds) *2017 IEEE Virtual Reality (VR) Conference Proceedings*, Los Angeles, CA, USA, 18–22 March 2017. pp. 45–54.

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.

Malone, S. & Brünken, R. (2016). *The Role of Ecological Validity in Hazard Perception Assessment*. Transportation Research Part F: Traffic Psychology and Behaviour, 40: 91–103.

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.

McKenna, F. P., Horswill, M. S. & Alexander, J. L. (2006). *Does Anticipation Training Affect Drivers' Risk Taking?* Journal of Experimental Psychology, 12(1): 1–10.

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.

Nosofsky, R. M. (1986). *Attention, Similarity, and the Identification-Categorization Relationship*. Journal of Experimental Psychology: General, 115(1): 39–61.

Pai Mangalore, G. (2019). *The Promise of VR Headsets: Validation of a virtual reality headset-based driving simulator for measuring drivers' hazard anticipation performance* [Dissertation]. Amherst, MA: University of Massachusetts Libraries.

Pammer, K. & Blink, C. (2013). *Attentional Differences in Driving Judgments for Country and City Scenes: Semantic congruency in inattentional blindness*. Accident Analysis & Prevention, 50: 955–963.

Poulsen, A. A., Horswill, M. S., Wetton, M. A., Hill, A. & Lim, S. M. (2010). *A Brief Office-Based Hazard Perception Intervention for Drivers with ADHD Symptoms*. Australian and New Zealand Journal of Psychiatry, 44(6): 528–534.

Pradhan, A. K. & Crundall, D. (2017). Hazard avoidance in young novice drivers: Definitions and a framework. In D. L. Fisher et al. (eds), *Handbook of Teen and Novice Drivers* (pp. 81–94). Boca Raton, FL: CRC Press.

Pradhan, A. K., Hammel, K. R., DeRamus, R., Pollatsek, A., Noyce, D. A. & Fisher, D. L. (2005). *Using Eye Movements to Evaluate Effects of Driver Age on Risk Perception in a Driving Simulator*. Human Factors, 47(4): 840–852.

Pradhan, A. K., Pollatsek, A., Knodler, M. & Fisher, D. L. (2009). *Can Younger Drivers Be Trained to Scan for Information that Will Reduce Their Risk in Roadway Traffic Scenarios that Are Hard to Identify as Hazardous?* Ergonomics, 52(6): 657–673.

Prothero, J. D. (1998). *The Role of Rest Frames in Vection, Presence and Motion Sickness* [PhD thesis]. Ann Arbor, MI: UMI.

Prothero, J. D. & Parker, D. E. (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* (pp. 47–66). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Rebenitsch, L. & Owen, C. (2016). *Review on Cybersickness in Applications and Visual Displays*. Virtual Reality, 20(2): 101–125.

Ruddle, R. A. & Lessels, S. (2009). *The Benefits of Using a Walking Interface to Navigate Virtual Environments*. ACM Transactions on Computer-Human Interaction, 16(1): 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(2): 157–168.

Saredakis, D., Szpak, A., Birckhead, B., Keage, H. A. D., 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.

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 & Prevention, 42(6): 1577–1584.

Simons-Morton, B. G., Guo, F., Klauer, S. G., Ehsani, J. P. & Pradhan, A. K. (2014). *Keep Your Eyes on the Road: Young driver crash risk increases according to duration of distraction*. Journal of Adolescent Health, 54(5): S61–S67.

Slater, M. & Sanchez-Vives, M. V. (2016). *Enhancing Our Lives with Immersive Virtual Reality*. Frontiers in Robotics and AI, 3: 74.

Thomas, F. D., Rilea, S., Blomberg, R. D., Peck, R. C. & Korbelak, K. T. (2016). *Evaluation of the Safety Benefits of the Risk Awareness and Perception Training Program for Novice Teen Drivers.* (Report no. DOT HS 812 235). Washington, DC: National Highway Traffic Safety Administration.

Trick, L. M. & Caird, J. (2011). Methodological issues when conducting research on older drivers. In D. L. Fisher et al. (eds), *Handbook of Driving Simulation for Engineering, Medicine and Psychology*. Boca Raton, FL: CRC Press. pp. 26-1 – 26-13.

Underwood, G. (2007). *Visual Attention and the Transition from Novice to Advanced Driver*. Ergonomics, 50(8): 1235–1249.

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.

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

Ventsislavova, P., Crundall, D., Baguley, T., Castro, C., Gugliotta, A., Garcia-Fernandez, P., Zhang, W., Ba, Y. & Li, Q. (2019). *A Comparison of Hazard Perception and Hazard Prediction Tests Across China, Spain and the UK*. Accident Analysis & Prevention, 122: 268–286.

Ventsislavova, P., Gugliotta, A., Peña-Suarez, E., Garcia-Fernandez, P., Eisman, E., Crundall, D. & Castro, C. (2016). *What Happens When Drivers Face Hazards on the Road?* Accident Analysis & Prevention, 91: 43–54.

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(1): 153–160.

Wallis, T. S. & 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.

Weidner, F., Hoesch, A., Poeschl, S. & Broll, W. (2017). *Comparing VR and Non-VR Driving Simulations: An experimental user study*. Paper presented at the *2017 IEEE Virtual Reality (VR) Conference,* Los Angeles, CA, USA, 18–22 March 2017. pp. 281–282.

Wells, P., Tong, S., Sexton, B., Grayson, G. & Jones, E. (2008). *Cohort II: A study of learner and new drivers*. Volume 1, Main Report. Road Safety Research Report. London: Department for Transport.

Wetton, M. A., Hill, A. & Horswill, M. S. (2013). *Are What Happens Next Exercises and Self-Generated Commentaries Useful Additions to Hazard Perception Training for Novice Drivers?* Accident Analysis & Prevention, 54: 57–66.

RAC
Foundation

Mobility • Safety • Economy • Environment

The Royal Automobile Club Foundation for Motoring Ltd is a transport policy and research organisation which explores the economic, mobility, safety and environmental issues relating to roads and their users. The Foundation publishes independent and authoritative research with which it promotes informed debate and advocates policy in the interest of the responsible motorist.

RAC Foundation
89–91 Pall Mall
London
SW1Y 5HS

Tel no: 020 7747 3445
www.racfoundation.org

Registered Charity No. 1002705

August 2021 © Copyright Royal Automobile Club Foundation for Motoring Ltd

Designed and printed by
The Javelin Partnership Ltd
Tel: 0118 907 3494

Produced on paper from a managed
sustainable source which is FSC certified
as containing 50% recycled waste.