A novel driving assessment combining hazard perception, hazard prediction and theory questions

David Crundall, Editha van Loon, Thom Baguley and Victoria Kroll

Department of Psychology

Nottingham Trent University

This is a postprint of Crundall, D., van Loon, E., Baguley, T., and Kroll, V. (2020). A novel driving assessment combining hazard perception, hazard prediction and theory questions. *Accident Analysis and Prevention, 149,* 105847. It was submitted to https://irep.ntu.ac.uk/ on 14/10/20.

Author Note

This research was funding by the UK Department of Transport through the Innovation Challenge Fund. We would like to thank Scott Fox, John Groeger, Guy Chamberlain and the DVSA for their assistance. We are grateful for the help of Pearson VUE for allowing access to their premises for data collection purposes and would also like to thank Jellylearn for creating our CGI stimuli.

Abstract

A new hazard test was created using high-fidelity computer animation containing ten hazards. Sixty learner drivers and sixty experienced drivers sat either a hazard-perception version of this test (requiring timed responses to materialized hazards) or a hazard-prediction variant of the test (where the screen is occluded as the hazard begins to appear and drivers are asked 'What happens next?'). Recent studies have demonstrated that the prediction test format outperforms the hazard perception format using naturalistic video, but there has not yet been a study replicating this effect with computer-animated materials similar to the quality of those used in the official UK hazard perception test. The new test also included eleven theory questions designed to probe drivers' knowledge of the rules of the road. The results demonstrated that both test variants differentiated between driver groups with considerable effect sizes. Theory-question scores were comparable across learner and experienced driver groups, reflecting learners' preparation for the test and possible issues with memory decay and overwriting in the experienced group. As an interesting aside, driving-related video game play negatively correlated with hazard perception performance, but not with hazard prediction scores. Some individual hazards better suited the prediction or perception test format, raising the possibility of a future hybrid test that combines the two approaches.

Development of a novel hazard test for drivers

Introduction

For nearly a decade, traffic-collision fatality rates in the UK have plateaued at around 1800 deaths per year. This unacceptable rate has proved difficult to reduce despite a raft of changes in legislation, training and assessment. One consistent pattern in the annual statistics is the over-representation of the youngest drivers in traffic collisions (e.g. Kinnear et al., 2013; Underwood, 2007). This pattern is not peculiar to the UK, but is noted across Europe (Adminaite et al., 2017), and many other countries, such as the US (IIHS, 2019), Oman (Al-Aamri, Padmadas, Zhang, & Al-Maniri, 2017), Iran (Moafian et al., 2013), and Malaysia (Ismail et al., 2016). Even the safest countries, such as Norway, do not escape this pattern of fatalities (Elvik, 2010).

One of the most successful interventions to target young-driver crash risk in the UK is the hazard perception test. Hazard perception skill is often considered akin to on-road situation awareness (Horswill and McKenna, 2004), and is defined as the ability to detect and recognize overt or developing hazards in the driving environment (McKnight & McKnight, 2003). This skill is typically assessed via a test that requires learners to watch a series of video clips from the point of view of a driver, and to press a button as soon as they see a developing hazard. The more quickly drivers respond to the hazards, the more points they receive, with higher scores reflecting greater hazard perception skill and subsequent on-road safety. Introduced in 2002, it has significantly reduced collisions in young drivers (Wells et al., 2008) with one estimate suggesting that the test prevents over 1000 injury-related collisions per year with an annual saving of £89.5 million (Horswill, 2016). Underlying this success are decades of research undertaken across the globe, with a majority consensus that hazard perception tests can measure underlying higher-order skills that are relevant to driving (for a review see Horswill, 2017).

Performance on such tests has even been noted to predict retrospective and prospective involvement in traffic collisions (e.g. Boufous et al, 2011; Horswill et al., 2015). Several countries now include a hazard perception test in their driver licensing process, with more countries considering it. The instigation of national tests leads to an understandable demand for training resources, and a niche, yet global, industry has sprung up to service this need. One recent study has shown that, when operationalised appropriately, hazard perception training can significantly reduce the future crash likelihood of young, male drivers (Thomas et al., 2016).

The story is not completely positive however, and several research groups have failed to create hazard tests that are sensitive to driver safety and experience (see Crundall, 2016; Moran et al., 2019). There are a number of potential reasons for mixed findings in the literature, not least the fact that there is no agreed method for developing a hazard test. Some tests developed for research purposes use static imagery instead of dynamic clips (Huestegge, et al., 2010; Scialfa, et al., 2012; Vlakveld 2014); some use video-recorded hazards (McKenna & Crick, 1994; Crundall et al., 2016) while others employ computer-generated imagery (e.g. Malone & Brünken, 2016); and some use naturalistic hazards (e.g. Horswill et al., 2008, 2015) while others stage dangerous events (e.g. Jackson et al., 2009; Borowsky et al., 2010). The recorded measures of hazard perception skill differ also. While, the traditional response is a speeded-button press to an unfolding dynamic hazard, a recent review of hazard perception methodologies (Moran et al., 2019) found 117 different measures of hazard perception across 48 studies. Such inconsistencies are perhaps inevitable in a research field where we do not even have a common vocabulary (Pradhan & Crundall, 2017).

In the context of this rapidly evolving field of research, it is important to ensure that any new developments in hazard perception assessment are documented thoroughly and, where possible, compared to existing formats to ascertain whether there is any benefit. The current paper discusses a new format for a hazard test and compares this to a more traditional method using behavioural data and self-reported preferences.

Hazard Perception vs Hazard Prediction

One relatively recent development in hazard perception assessment is the rise of a new measure: hazard prediction. This method has evolved from early versions (Jackson, Chapman and Crundall, 2009; McGowan and Banbury, 2004; Vogel et al., 2003) based on the Situation Awareness Global Assessment Technique (Endsley, 2000), into a more refined challenger to the traditional hazard perception methodology (Crundall, 2016; Crundall & Kroll, 2018; Ventsislavova et al., 2019). The basic premise behind hazard prediction is that the safest drivers do not wait for a hazard to happen and then respond, but instead try to predict what will happen based on clues in the visual scene (i.e. hazardous precursors; Pradhan and Crundall, 2017). The effectiveness of such predictions is assessed by pausing the test (occluding the screen just as the hazard begins to materialise) and probing participants' understanding of how the scene might unfold. Early versions did just this: playback was interrupted and participants were asked, "What happens next?" Verbal or written answers were then scored for accuracy (e.g. Jackson et al., 2009). Later studies iterated the methodology, with several refinements including the provision of multiple-choice options following occlusion of the video. Faced with four such options, participants merely have to press a button to indicate their choice (e.g. Ventsislavova & Crundall, 2018; Kroll et al., 2020). This particular development simplified and automated the scoring process, while ensuring that drivers considered a range of potential outcomes when deciding on their answer.

Studies using hazard prediction tests have consistently found them to differentiate between groups of drivers based on safety or experience (e.g. Jackson et al., 2009; Castro et al., 2014; Crundall, 2016). Most recently, Horswill, Hill and Jackson (2020) found scores on a variant of the hazard prediction test to relate to self-reported collisions. This test involved recording participants' verbal predictions following occlusion of the clip, which were then scored according to a scoresheet based on driving experts' predictions. Furthermore, in two studies where hazard prediction clips were pitted against hazard perception clips (where both sets of clips were identical save for the early occlusion in the prediction clips), the prediction test was more successful in differentiating driver groups (Crundall & Kroll, 2018; Ventsislavova et al., 2019).

All of the hazard prediction studies above have used video-recorded hazards. However, the official UK hazard perception test transitioned to the use of videos comprised of computer-generated imagery (CGI) in 2015. If hazard prediction is to be considered as a serious contender for a national test, we must consider whether the effects noted in video-based hazard tests translate into the medium of CGI.

While there have been a number of studies that have used simulators to assess hazard avoidance, very few studies have used CGI clips, especially where hazard prediction is the main task. Vlakveld (2014) provides one of the few exceptions. He compared learner drivers and professional drivers on two CGI hazard tests that required participants to identify hazards that *might have occurred* given the circumstances. While not strictly a hazard prediction test (as there were only *potential hazards* to predict rather than actual ones), he found professional drivers to outperform learners on both tests. Malone and Brünken (2015) have also employed CGI animations in their hazard test. In one variant, they required participants to choose from four options following the hazard clip, and found a significant difference in performance between learners and experienced drivers. Yet again, this variant was not actually a hazard prediction test, as half of the target hazards were merely potential (and would never appear), while the other half fully materialised during the clip before the probe question. In a larger subsequent study, Malone and Brünken (2016) again found multiple-choice options to differentiate between driver groups, though the experience-novice performance gap was greater when drivers were asked to make a speeded-button response (i.e. the more traditional hazard perception measure was the most effective). These data are in contrast to those found by Crundall and Kroll (2018) and Ventsislavova et al. (2019) where hazard prediction tests better differentiated driver groups than hazard perception tests.

Though comparisons across these studies are confounded by the differences in the exact methods employed, the possibility remains that the success of the hazard prediction test as noted in many studies using naturalistic video recordings, may not translate into the CGI world. One possible reason for this is that hazard perception relies on behavioural and environmental subtleties, which may not be captured by animators (such as the gaze direction of a pedestrian thinking of stepping into the road). The current study provided an opportunity to test this hypothesis, pitting a hazard perception test against a hazard prediction test using high-fidelity CGI hazard clips.

A novel hazard test

The current study included two additional innovations beyond the use of hazard prediction clips. First, we aimed to combine our hazard test with driving theory questions. In the UK, learner drivers must pass both a hazard perception test and a multiple-choice theory test (including questions on appropriate behaviour in certain situations, understanding of road signs, etc.). Unfortunately, the current theory test presents questions often devoid of context[1], removing real-world cues to stored knowledge, but also removing the additional driving demands that may interfere with knowledge recall while on the road. A more realistic test of applying theory knowledge needs greater context. We argue that drivers should be trained to use theory knowledge in context, and while under additional cognitive demands from concurrent driving-related tasks. This may improve the usefulness and ecological validity of theory-related questions.

Adding theory questions to a hazard clip may also have benefits for the hazard assessment. Current hazard tests lack the additional cognitive load of concurrent driving-related tasks (though see Isler, 2009, for one of the few exceptions). While avoiding hazards should be the primary goal of driving, it is not the only task to demand our attention on real roads. Testing hazard perception skill in the absence of secondary demands (e.g. navigation decisions) may over-estimate learner and novice driver skills. The inclusion of theory questions into hazard clips, inserted at meaningful but unexpected points, increases the cognitive load of the overall task and should stop participants thinking solely about the location of the next hazard (which is admittedly a desirable trait in a driver, but is unrealistic in real-world settings).

A second innovation is the use of one long CGI clip that lasts 10 minutes (containing 10 hazards), rather than the traditional approach that uses multiple short clips, often less than 60 seconds in length. While many studies of vigilance decrements use long-duration tasks of one or more hours (Gartenberg, Gunzelmann, Hassanzadeh-Behbaha & Trafton, 2018), others have noted significant decrements within 30 minutes of a monotonous task (Pattyn, Neyt, Henderickx,

& Soetens, 2008). In the field of hazard prediction, decrements have been found with much shorter instances of time-on-task (24s vs 44s; Crundall, 2016).

The findings of Crundall (2016) notwithstanding, the traditional use of multiple short clips likely underestimates the true decline in vigilance for hazards in the real world. Such short clips may artificially encourage maintenance of attention to the task, especially where the context of one clip differs greatly from the preceding clip (e.g. a sudden transition from the motorway to a rural road may reignite interest in the task). Cognitive control theories suggest that vigilance (performance) decrements over time occur due to an increased chance of losing cognitive control over the attentional set (i.e. the longer the task, the more likely the mind-wanders from the main goal). In such cases, multiple, short clips give greater opportunity to reset cognitive control (Ariga and Lleras, 2011), and may therefore over-estimate performance compared to real driving (see Figure 1).

The use of a single long clip with multiple hazards is impossible to achieve through video-recordings of everyday driving as naturally occurring hazards occur relatively infrequently. It is easier to use a single long clip where no particularly hazardous events occur, and one asks the viewer "What *might* occur…?"), but we argue that this does not reflect true hazard prediction. The rationale for this is that it is relatively easy to note that a car in a side road *may* pull out in front of you, but it is more difficult to identify whether the car *will* pull out, or whether it will wait until you have passed. This is the crux of hazard prediction.

Recently, we have been able to collect footage from emergency response vehicles that contained multiple hazardous events within a single clip, though this was dependent on the fact that the drivers were on blue-light training runs (Kroll et al., 2020). When filming from ordinary vehicles, our natural hazards are much less frequent. Fortunately, CGI offers the opportunity to create multiple hazards within the same clip, thus mitigating the potential for sudden changes in environment to reset cognitive control.

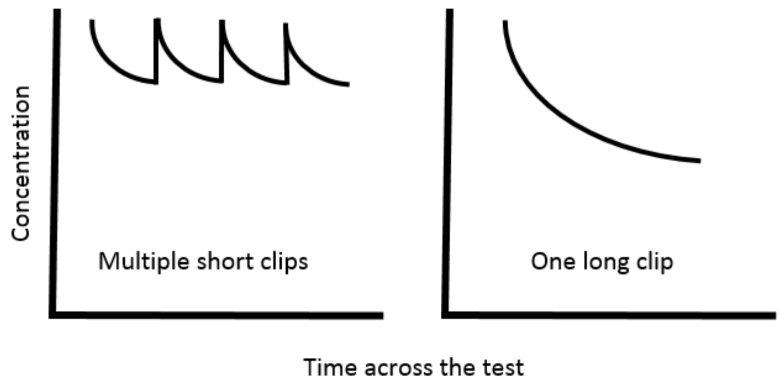

Figure 1. The hypothetical decrease in concentration on the task with multiple short hazard clips and one long hazard clip

Alas, with three innovations (hazard prediction, combining theory and hazard assessment, and the use of one long hazard clip rather than multiple short clips) it was not practical to manipulate all of them experimentally. As the comparison of the hazard prediction and hazard perception test formats was considered most important, the study was designed to compare these directly in a between-groups design, with a group of experience drivers and learner drivers undertaking a perception test, and another group of experienced and learner drivers undertaking the hazard prediction variant. Although manipulation of the other innovations was unfortunately impractical, in lieu of performance data, we sought drivers' opinions on our additional innovations compared to the current standard in UK hazard perception testing.

Is hazard perception skill influenced by video game play?

An additional research question concerned the impact of video game playing on hazard assessment scores. There are many aspects of the current UK hazard perception test that could be argued to overlap with video games, including the use of CGI, the adoption of the driver's perspective, and the requirement to make time-critical responses. Some driving instructors have informally commented to the lead author that the hazard perception test is too similar to a video game, and therefore likely favours those learners who play such driving games. There is certainly evidence in the literature that playing video games can improve a variety of perceptual and attentional processes, cognitive control and fine motor-skills (Achtman, Green, & Bavelier, 2008; Dye, Green, & Bavelier, 2009; Green & Bavelier, 2003, 2007; Howard, et al., 2016). In regard to driving, Rupp et al. (2016) found that video game players produce better lane maintenance in a distracted driving task, while Vlakveld (2014) found that those who report greater video-game playing performed better on his hazard perception tests. If video-game playing genuinely improves hazard perception skill on real roads, then this is a positive effect that is reflected in test performance. However, it is also a possibility that the game-players are merely better, or more comfortable, at interacting with computer-based, first-person assessments. If this is the case, the hazard perception test may over-estimate game-players' hazard detection abilities.

In contrast, other studies have failed to find effects of video game play on drivers' eye movements when viewing road scenes (Ciceri and Ruscio, 2014), or when driving on real roads (Wayne and Miller, 2018). Some studies have even noted a negative relationship between video game play and risk-taking in driving scenarios (Achtman et al., 2008; Deng et al., 2017). The current study provided an opportunity to relate hazard performance to self-reported video game play.

The current study

To assess the impact of our innovations, we commissioned a ten-minute CGI clip from the same company that produces the high-fidelity CGI clips for the official UK test. The clip contained 10 hazards designed by our team of traffic psychologists. We added multiple-choice theory probe questions at relevant points during the clip. The theory questions were similar to those encountered in the official UK theory test, and were chosen with the assistance of an expert from the UK Driver and Vehicle Standards Agency. The resultant clip provided the material for the hazard perception test, though the hazard prediction test required further editing: Occlusion points for the ten hazards were selected and multiple-choice options were added post-occlusion.

Following the success of the prediction test format over the more traditional hazard perception format in previous video-based studies (Crundall & Kroll, 2018; Ventsislavova et al., 2019), we predicted that the dominance of the prediction test would be found in our CGI clips. However, we also acknowledged the possibility that Malone and Brünken's (2016) findings could presage the opposite results with CGI media.

Method

Participants. One hundred and twenty participants were recruited for this study, with 60 classified as learner and novice drivers (34 Female, average age 23.3 years, SD 7.3) and 60 classified as experienced drivers (33 Female, average age 39.6, SD 10.8). Of the learner and novice drivers, 56 had taken the hazard perception test at least once. Five of the drivers had taken the on-road test, of which two had passed (both within the last 6 months, hereafter included in the category simply referred to as 'Learners'). Learner drivers reported to be driving an average of 1.8 hours a week, though several participants left this field blank in the demographics questionnaire. Experienced drivers averaged 8.3 hours of driving per week, and reported an average mileage of 7618 miles a year. A minimum of three years of post-licensure driving was required. Twenty of the experienced drivers (33%) were young enough to have taken the hazard perception test as part of their theory test. See Table 1 for a more detailed breakdown of participant demographics across their assigned conditions.

Table 1. Participant demographics across assigned conditions.

	Learners		Experienced	
	Perception Test	Prediction Test	Perception Test	Prediction Test
Total *N*	30	30	30	30
Females (*N*)	19	15	18	15
Age in year (*SD*)	22.6 (6.8)	23.9 (7.3)	39.3 (10.9)	40.0 (10.8)
Mean years since passing test	-	-	19.0 (10.8)	21.4 (10.9)
Taken on road test? (*N*)	2	3	60	60
Mean attempts at on road test	-	-	1.7	1.8
Passed on-road test? *(N)*	1	1	60	60
Taken Hazard Perception test (*N*)	29	27	10	10
Passed Hazard Perception test (*N*)	22	23	10	10
Mean attempts at HP test	1.8	1.6	-	-
Mean most recent score on HP test (SD)	57.5 (7.4)	55.0 (10.0)	-	-
Passed Theory test? (*N*)	19	18	30	30

Mean most recent score on Theory test (SD)	44.0 (4.1)	43.9 (3.4)	-	-
Mean total hours of practice with instructor	18.2	25.4	-	-
Mean total hours of practice with family/friends	8.8	5.4	-	-
Mean hours driving per week	2.1	1.5	7.7	8.9
Mean annual mileage	-	-	8072	7179

Design. A 2 x 2 between-groups design compared driving experience (experienced and learner drivers) across test-variant (with half of the learners and experienced drivers assigned to the *hazard perception test*, while the remaining participants were assigned to the *hazard prediction test*). The dependent variables for the hazard perception test were the response times to detect a hazard, and accuracy for detecting hazards (i.e. making a response within the temporal scoring window). Response times were converted into scores following the method used for the national UK test: the scoring window (from hazard onset to hazard offset) is divided into 5 equal segments, with responses in the earliest segment scoring five points, and decreasing scores given to responses in later segments. Hazard onset is defined as the point at which the hazard begins to develop (e.g. the car ahead indicates and begins to change lanes in Hazard 2; the motorcycle is first visible in Hazard 3). Note that some clues or precursors to the hazard are visible prior to hazard onset (the presence of the occluding HGV in Hazard 3). Hazard offset is defined as the point at which the hazard is no longer dangerous, or a collision would have already occurred without a response. In some cases this was coincident with the disappearance of the hazardous object (e.g. the motorcycle in Hazard 3 quickly passes in front of the participant's vehicle and disappears off-screen), while in others the hazardous object remain on-screen but was no longer considered hazardous. For example, in Hazard 2 an overtaking car pulls into the participant's lane abruptly and then speeds off. Once it has accelerated away it is no longer hazard, even though it is still visible on the screen. The number of hazards that receive a response within these scoring window determines overall accuracy.

The primary dependent variable for the hazard prediction task is the number of hazards that participants correctly identify when asked, “What happens next?” As the prediction test does not collect response times, we cannot compare these tests directly, but we can analyse the percentage of hazards detected in both conditions (i.e. the percentage of hazards that received a response in the perception test with the percentage of correct answers given in the prediction test; Crundall and Kroll, 2018; Ventsislavova et al., 2019). Scores on the theory test were also compared directly across the two test-variants, though we had no reason to believe that these would differ.

Stimuli and Apparatus. Both the hazard perception test and the hazard prediction test were identical in terms of Computer-Generated Imagery (CGI) content. This content was a 10-minute continuous drive through a CGI-rendered world, across a variety of roads (arterial, suburban and rural). The video took the perspective of a driver, travelling through junctions, turning into side roads, and encountering 10 pre-specified hazards (see Table 2 for a brief description of each hazard, and Figure 2 for an example screen shot).

When undergoing the hazard perception test, the 10-minute clip played in full, with the hazards fully materialising. The only interruptions in playback were for 11 theory questions that would appear at appropriate points within the clip. The hazard prediction test also included the 11 theory questions, with an additional 10 “what happens next?” (WHN) questions replacing the need for speeded responses to hazards. These WHN questions were presented on the screen following occlusion of the clip just after each hazard had begun to materialise (i.e. if participants were looking in the right place at the right time, they would have seen several frames containing the imminent hazard). In some cases, the occlusion point was almost identical with the hazard onsets used to create the scoring windows for the hazard perception test (e.g. Hazard 3), while in others the occlusion point followed the hazard onset by a second or more. For example, in Hazard 2, the onset is defined by a silver car ahead beginning to change lanes. The actual hazard is the subsequent behaviour of the red car in the lane to your left, who decides to suddenly overtake the silver car. The occlusion point in this clip occurs as the red car beings to make his manoeuvre. WHN questions were accompanied by four text options to choose between. Participants pressed a corresponding button on the keyboard (1 to 4) to register their response. After a response was given, the clip would resume from the point at which it occluded.

Figure 2. A screen shot from hazard 3: As you begin to turn right at a junction, an oncoming motorcycle becomes visible that was previously occluded by a turning HGV. The ellipse

identifies where the motorcycle is for the purpose of this figure, but this did not appear in the actual clip.

The eleven theory questions, presented in both the hazard perception and hazard prediction tests, were given in a similar format to the WHN questions, with four options to choose between. A corresponding keyboard response was required before the test would resume. A list of the theory questions can be found in Table 3. Both tests were silent apart from a voice-over of an instructor providing guidance on where the film car would turn (e.g. “take the next left”).

Table 2. A brief description of the hazards in the current test, their onset times for the hazard perception test, and the multiple-choice answers for the hazard prediction test. Where descriptions include "You intend to…" these driver intentions were imparted to participants via a voice-over, akin to a driving instructor telling the driver where to turn next.

Hazard No.	Description	Hazard window duration (ms)	Multiple-choice options (Correct answer is underlined)
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.	4400	• The parked blue car on the left indicates and pulls off as you pass it. • The oncoming car turns into a side road, but must stop, blocking your way. • A white van pulls out of the side road on the left, forcing you to brake. • The oncoming car accelerates towards you, preventing you from overtaking the parked car ahead. •
2	While travelling in the right lane of a two-lane carriageway, the car immediately ahead, indicates and moves over into the left lane. Unfortunately, he fails to see a car in the left lane, hidden in his 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.	4000	• The red car in the left lane suddenly pulls into your lane. • The oncoming car turns sharply across your path in order to enter a driveway on your left. • The silver car ahead suddenly swerves back into your lane. • The silver car brakes harshly, forcing you to brake also.
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.	2480	• The LGV decides not to turn right, and proceeds straight across the junction narrowly missing you. • A pedestrian steps into the road that you are trying to turn into. • An oncoming motorcycle prevents you from turning. • There is congestion on the road you are turning into, which forces you to stop.
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	2000	• The passenger door of a car parked on the right suddenly opens.

	parked vehicles. A second driver, visibly approaching from a side-road, misinterprets this signal to suggest he can pull out. As you drive forward, the car suddenly emerges from the side road.		• A car emerges from a side street on the right, into your path. • A pedestrian steps into the road from between parked cars on the left. • The red car parked on the left, indicates and pulls off in front of you.
5	While 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.	2000	• The white parked car on the left tries to pull off as you pass it. • A man carrying a large box steps out from behind a white van parked on the right. • An oncoming car turns across your path to enter a driveway on your left. • A woman pushing a buggy steps out from between parked cars on the left.
6	While driving along a suburban route, a cyclist can be seen on a cycle lane shared with the pavement on the left. He is travelling in the same direction as you, but you quickly pass him. 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.	1000	• A pedestrian steps into the side road as you begin to turn. • As you attempt to turn, a car from right passes you heading for the same side road. • As you turn into the side road you find immediate congestion ahead that forces you to brake. • A cyclist crosses the side road as you begin to turn.
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.	2400	• A pedestrian runs into the road from the left from behind a parked car. • The lights at the pedestrian crossing turn red forcing you to stop. • The blue car parked on the left suddenly indicates and tries to pull off in front of you. • The car ahead suddenly brakes forcing you to brake also.
8	When 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.	3000	• The bus indicates and starts to pull off as you attempt to pass it. • A pedestrian emerges from in front of the bus on the left. • A car emerges from a side road on your left.

			• The oncoming car accelerates towards you, preventing you from overtaking the bus.
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.	4240	• The oncoming car strays into your lane. • A car enters the mini-roundabout ahead from the right. • A pedestrian crosses the road from the right. • A car enters the mini-roundabout from the left.
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.	2720	• An oncoming motorcycle prevents you from turning. • One of the cars waiting in the oncoming lane closes the gap into the side road, preventing you from turning. • A red car emerges from the side road on the right and pulls out in front of you. • A pedestrian steps out from between waiting cars on the right.

Table 3. The theory questions, with multiple-choice options, used in the current test

Question No.	Context	Question and multiple-choice options (Correct answer is underlined)
1	You are approaching a T-junction. A high wall shields the view to the right.	Why must you take great care when turning at this junction? • The road surface is poor • The footpaths are narrow • The road markings are faint • The view is restricted
2	You are passing a fire station with a yellow-hatched box in front.	When may you enter a box junction? • When there are fewer than two vehicles ahead • When signalled by another road user • When your exit road is clear • When traffic signs direct you

3	You are driving along a residential road.	There are no speed-limit signs on the road. How is a 30 mph limit indicated? • By hazard warning lines • <u>By street lighting</u> • By pedestrian islands • By double or single yellow lines
4	You have just encountered a hazard cause by an ambiguous headlight signal (see Hazard 4).	Other drivers may sometimes flash their headlights at you. According to the Highway Code, what does this signal mean? • There's a radar speed trap ahead • They're giving way to you • <u>They're warning you of their presence</u> • There's a fault with your vehicle
5	You have just encountered an emergency vehicle on flashing blue lights.	What should you do if you're being followed by an ambulance showing flashing blue lights? • <u>Pull over as soon as it's safe to do so</u> • Accelerate hard to get away from it • Maintain your speed and course • Brake harshly and stop well out into the road
6	You have just passed through a flashing amber light at a pelican crossing.	What must you do when the amber light is flashing at a pelican crossing? • Stop and wait for the green light • Stop and wait for the red light • Give way to pedestrians waiting to cross • <u>Give way to pedestrians already on the crossing</u>
7	You are driving on a road that is becoming increasing rural.	It rains after a long, dry, hot spell. How can this affect the road surface? • <u>It can become unusually slippery</u> • It can give better grip • It can become covered in grit • It can melt and break up
8	You are driving on a rural road where the central line markers change in length.	What do the long white lines along the centre of the road mean? • Bus lane • <u>Hazard warning</u> • Give way

		• Lane marking
9	You are driving on a rural road.	You are driving on a busy main road. What should you do if you find that you are travelling in the wrong direction? Turn into a side road on the right and reverse into the main road Make a U-turn in the main road Drive on to the next mini-roundabout and make a U-turn • <u>Turn around in a side road</u>
10	You pass a national speed limit sign.	What is the national speed limit on a single carriageway road for cars and motorcycles? • 30 mph • 50 mph • <u>60 mph</u> • 70 mph
11	You enter roadworks with temporary traffic lights, and a temporary 30 mph sign.	What must you do when entering roadworks where a temporary speed limit is displayed? • <u>Obey the speed limit</u> • Obey the limit, but only during rush hour • Ignore the displayed limit Use your own judgment; the limit is only advisory

Apparatus. The test was presented on a Lenovo laptop attached to a 28 x 16 cm screen, positioned at approximately 50 cm from the participant, creating a visual angle of 31 by 18 degrees. Participants listened to the directional voiceover via headphones. They responded to hazards and questions using the mouse and keyboard of the laptop, respectively.

Procedure. Participants were tested either in a laboratory or in a Government driving-test centre. Written instructions were provided and participants were required to sign a consent form, which detailed their rights to withdraw at any point without explanation, and to withdraw their data from the study at a later point. They were also asked to complete a demographic questionnaire. This included questions regarding driving history (whether they had passed the test, what they scored on the official test, hours spent learning to drive, etc.). One additional question asked drivers to rate their familiarity with playing driving-related video games on a scale from 1 to 5.

Participants then underwent either the hazard perception test or the hazard prediction test. Upon completion of the test, participants were given an evaluation questionnaire to assess their thoughts about the test they had just done. This questionnaire asked a series of questions including “Do you prefer video clips or CGI?”, “Do you prefer single hazard/theory tests or a combined test?”, and “Was your overall experience better or worse than that of the official HP test?” All evaluation questions were rated on a scale from 1 to 7. Following the questionnaire, all participants received a £10 Amazon voucher for taking part in the study.

Instructions for participants in the hazard perception condition included the following: “This clip will contain several hazards. Press the left mouse button when you observe a developing hazard. **A developing hazard is something that causes you to take action, like**

changing speed or direction to avoid a collision. Make sure you only press once the hazard starts to occur, but try to be as quick as possible after this point. Try not to press too many times to things that do not become a hazard, as we may not be able to use your data."

In the hazard prediction condition, the instructions were changed to, "This clip will contain several developing hazards. **A developing hazard is something that causes you to take action, like changing speed or direction to avoid a collision.** Before the hazard occurs, the clip will freeze and you will be presented with the question: what happens next? There will be four plausible options given on the screen, one of which is correct. The clip will provide sufficient information to indicate what the correct answer is, providing you were looking in the right place at the right time. There is no time limit associated with the question, but try to be as quick as possible. You only get one attempt at each hazard and are not able to review or change your responses once they are made." The instructions for both conditions also explained that theory questions would appear at frequent intervals, and described how to respond to them.

Results

Analyses are presented separately for the hazard perception test and the hazard prediction test, before the results of the two assessments are compared. Responses to questions regarding participant preferences for our innovations compared to the standard UK hazard perception test are reported at the end of the results section.

Hazard perception

Before analysing responses to hazards, the total number of responses for each participant were calculated. This measure included multiple clicks within the hazard windows and any clicks outside the hazard windows. The average number of clicks was 33 (3.3 per hazard) though two learners clicked 84 and 88 times, respectively. Both participants' responses frequencies were greater than three standard deviations above the mean of the sample, and were removed as outliers from the subsequent analysis of response times. Removing the two outliers reduced the overall sample mean to 31.2 clicks (3.12 per hazard), and the means of the groups became almost identical (31.6 and 30.8 for learners and experienced drivers, respectively). There was no evidence for a difference in the number of clicks made by the learners and experienced drivers (t_{56} = .26, p = .7).

The first response that a participant made which fell within a scoring window was awarded a score between five and one points, following the system used in the official UK HP test. Cronbach's alpha (which can be interpreted as a lower build of the true internal consistency) gave a score of 0.59, while Omega Categorical produced a score of .68, both suggesting room for improvement in internal consistency. These scores were converted into percentages of the total score possible. A comparison of the learner and experienced drivers revealed a clear difference between their respective scores of 34% and 49%, t_{56} = 4.55, p < .001, *Cohen's d* = 1.19.

Inspection of the group scores across the individual clips revealed that some clips produced larger group differences than others (Figure 3). When subjected to Bonferroni-corrected comparisons only clips 3, 5 and 9 remained significant. Despite this, there was sufficient difference between the groups across a number of clips to produce the overall significant effect when combined.

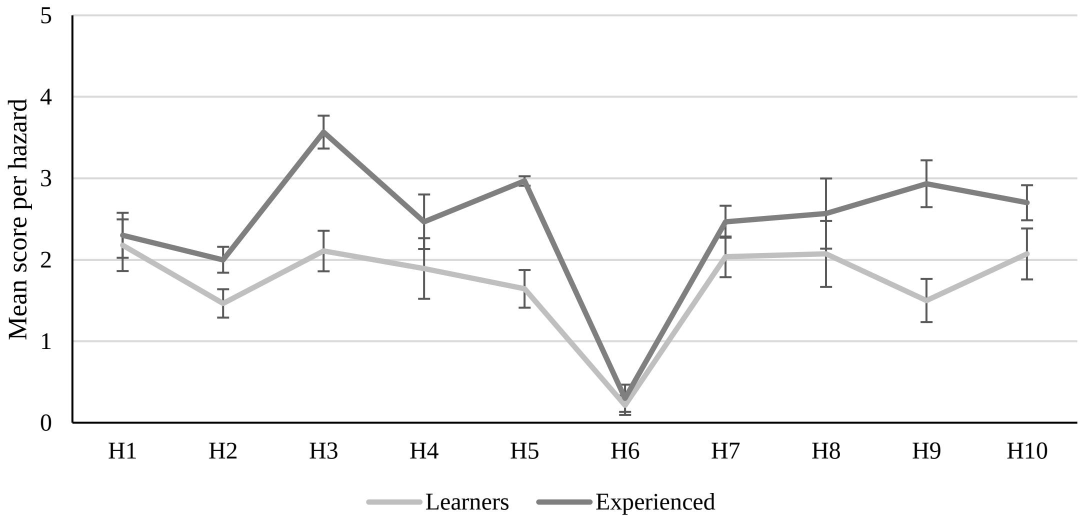

Figure 3. The mean score of learner and experienced participants across the 10 hazards. Hazards 3, 5 and 9 produced significant differences that withstood correction for familywise error (p < .005; with standard error bars added).

In addition to calculating HP scores, it is possible to categorise a response within a scoring window as an accurate detection of a hazard. Conversely, a lack of response within the window reflects a failure to detect the hazard. While this is a less-sensitive analysis of participant performance, it allows a more direct comparison with the hazard prediction performance that will be discussed shortly. Learners were found to respond to only 66% of the hazards within the scoring window, whereas experienced drivers responded to 78%, t56 = 2.5, p = .015, *Cohen's* d = .65.

Given the variation across hazards noted in Figure 3 (and the variation within the participant groups in terms of their driving experience), hits for all hazards in the hazard perception test were also analysed using a multilevel logistic regression with participants and hazards as random factors, and experience as a between-groups fixed effect. An intercept only model (with no predictor) estimated the *SD* of the participant random effect as 1.16 and the *SD* of the hazard random effect as 1.41, indicating that variance is split 41% and 59%, respectively. The deviance (likelihood ratio Chi Square, G2) for the intercept only model was 546. This decreased to 540.3 when the main effect of driver experience was added. This improvement in model fit was statistically significant, ΔG2(1) = 5.7, $p = .017$. This confirms that experienced drivers respond to more hazards on this test when variance between hazards (and participants) is accounted for, with estimated means of 86% [95% CIs: 68, 94] and 71% [95% CIs: 47, 87] respectively.

Hazard prediction

The hazard prediction test is much simpler to score than the hazard perception test. Participants do not make multiple timed responses removing the problem of outlying participants who might click too often. Instead, we simply calculate the percentage of hazards correctly predicted for learners and experienced drivers: 45% vs. 55%, respectively. This difference was significant, $t_{58} = 2.7$, $p = .008$, *Cohen's d* = .70.

The prediction data were also subjected to a multilevel logistic regression with participants and hazards as random factors, and experience as a between-groups fixed effect. The intercept only model estimated the *SD* of the participant random effect as 0.27 and the *SD* of the hazard random effect as 1.11, indicating that only 6% of the variation at level 2 of the model is attributable to participants, with hazards accounting for the majority of variance (94%). This raises the possibility that our simple t-test (which considers variation between hazards to be zero) may underestimate the associated standard errors. The deviance for the intercept only model was 716.1, which dropped to 708.9 following the addition of driver experience. This improvement in model fit was statistically significant, $\Delta G2(1) = 7.26$, $p = .007$. This confirms that experienced drivers perform better on this test when variance between hazards (and participants) is accounted for, with estimated means of 57% [95% CIs: 39, 73] and 44% [95% CIs: 27, 62] respectively.

Comparing hazard perception and hazard prediction

The impact of test-variant on performance was assessed by combining hazard perception accuracy and hazard prediction accuracy within a single multilevel logistic regression. An intercept only model estimated the *SD* of the participant random effect as 0.66 and the *SD* of the hazard random effect as 0.40, indicating that variance is split 62% and 48% between participants and hazards. The deviance for the intercept only model was 1530.3. Adding the main effects to the model reduced this to 1477.0, which was statistically significant ($\Delta G^2(2) = 53.2$, $p < .001$). When the interaction was included, it was not found to further reduce the deviance ($\Delta G^2(1) = .59$, $p = .44$). Dropping each main effect from the analysis and comparing back to the main effects model revealed both main effects to be significant ($G^2(1)_{testvariant} = 43.4$, $p <.001$, $G^2(1)_{drivergroup} = 13.1$, $p < .001$). This confirms that experienced drivers perform better than learners on both tests after accounting for variance due to differences between hazards and participants, with overall estimated means of 68.6% [95% CIs: 61, 75] and 56.1% [95% CIs: 48, 64], respectively.

There is no evidence that one test better differentiates between these groups than the other, though it is clear that accuracy on the hazard perception test is significantly higher than on the hazard prediction test, with estimated means across all participants of 73.4% [95% CIs: 67, 79] for the hazard perception test, and 50.3% [95% CIs: 43, 58] for the hazard prediction test (see Figure 4). This difference is understandable given that the hazards fully materialise in the hazard perception test, whereas drivers only get to see the full hazard in the prediction test *after* they have given their answer. This was especially problematic for hazards 1, 2, 3, and 5, where the precursors did not perhaps provide sufficient information to allow a correct prediction for the majority of participants. It is interesting to note however that the fully materialised hazards of the perception test did not produce 100% accuracy rates across all participants. Indeed, only three participants responded successfully to all 10 hazards in the perception test (and two of these were learner drivers).

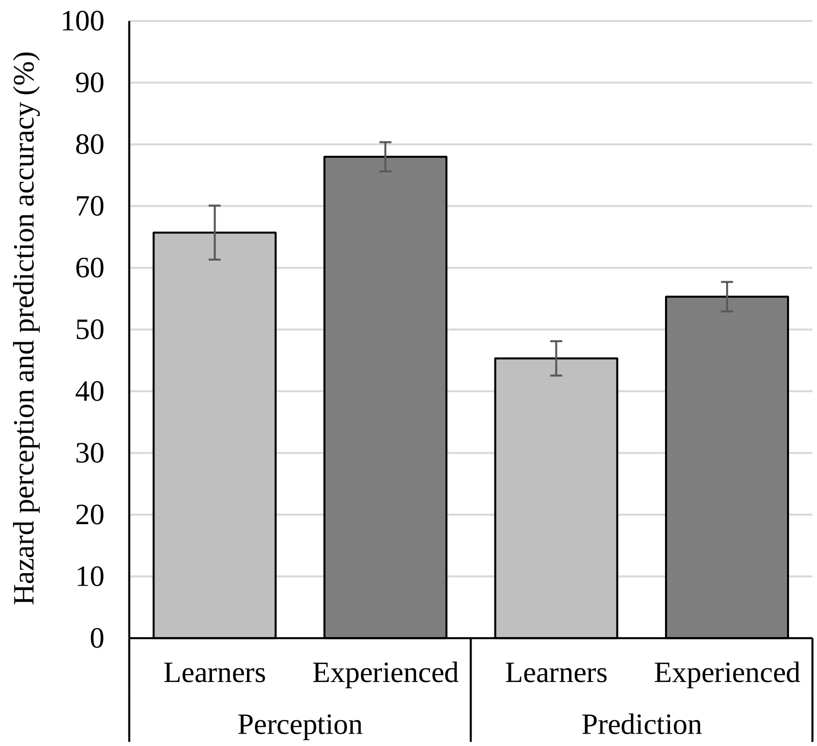

Figure 4. A comparison of accuracy across the two tests for all driver groups (with standard error bars added).

When comparing the tests across the individual hazards, similar differences are found between groups for certain hazards. Figure 5 reveals that hazard 3 and hazard 9 show experiential superiority in both tests, suggesting that participants' ability to respond within the time window is in part dictated by their ability to predict upcoming events. Likewise, several hazards show little sensitivity across the groups for both tests (e.g. hazards 1, 4, 8, 10). There are two hazards that stand out, however, for producing different patterns across the two studies. Hazards 5 and 6 suggest sensitivity on one test but not the other. H5 was very difficult to predict for all participants (a pedestrian emerging from between parked cars), but once the hazard triggered, the experienced drivers were much faster than the learners to respond. Conversely, H6 (a cyclist from behind, crosses the entrance to a side road) received very few responses within the time window from both groups, though the experienced drivers were much better at predicting the outcome in the prediction test.

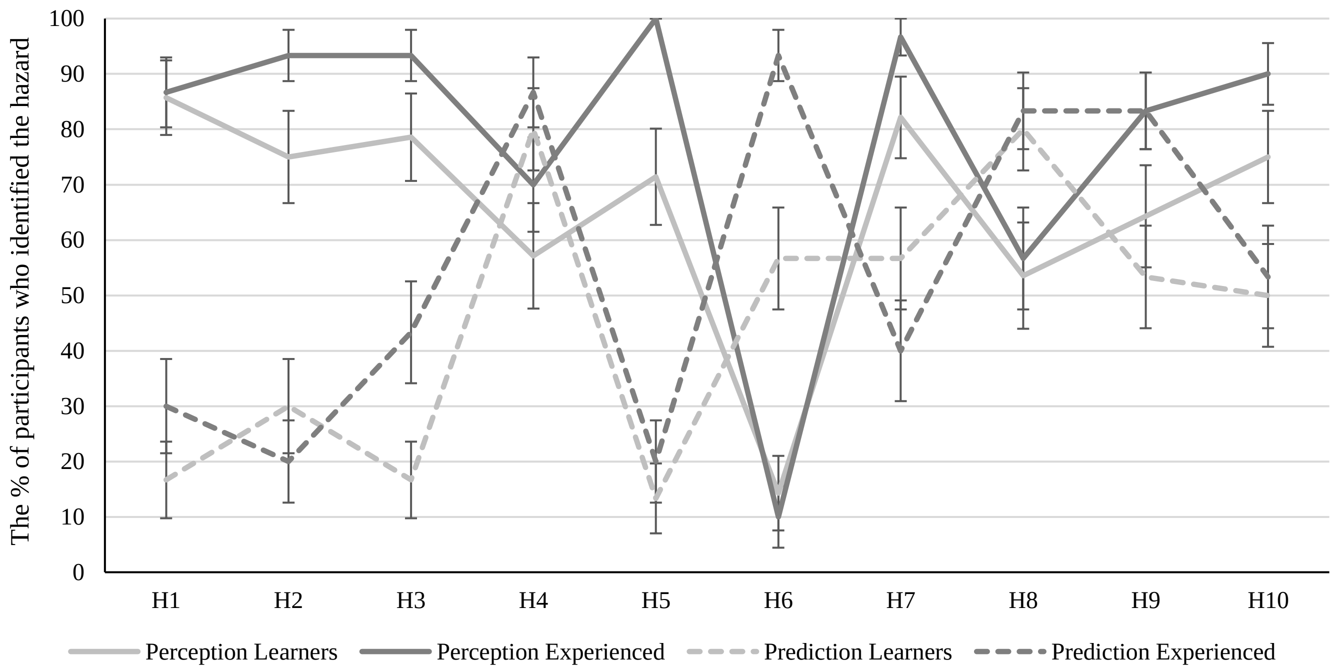

Figure 5. A comparison of accuracy, for individual clips, across the two tests for all driver groups (with standard error bars added).

Performance on theory questions

Participants' accuracy for individual theory questions were entered into a multilevel logistic regression, with participants and questions as random effects. An intercept only model estimated the *SD* of the participants as 0.77 and the *SD* of the question random effect as 1.60, with a 32% and 68% split of variance, respectively. The deviance for the intercept only model was 1132.8. The main effects model did not reduce the deviance (ΔG2(2) = 1.26, p = .53): Neither driver group ($G^2(1)$ = 1.24, p =.27) or test variant ($G^2(1)$ = .02, p =.90) were significant. The interaction model also did not reduce the deviance significantly (ΔG2(1) = 0.18, p = .67) and did not account for any variance beyond the intercept-only model. The estimated means for experienced drivers were 88.5% [95% CIs: 73, 96] and 87.9% [95% CIs: 72, 95] for the perception and prediction tests, respectively, while the learner drivers' estimated means were 84.8% [95% CIs: 66, 94] and 86.2% [95% CIs: 69, 95] for hazard perception and hazard prediction[2].

On several questions, experienced drivers performed surprisingly poorly. For instance, 26 out of 60 experienced drivers failed to answer question 4 correctly ("Other drivers may sometimes flash their headlights at you. According to the Highway Code, what does this signal mean?"). Of these 26 incorrect responses, 24 participants reported that the signal means, "They're giving way to you".

Equally worrying was that 20 out of 60 experienced drivers did not understand the meaning of a flashing amber light at a Pelican crossing (Q6). However, experienced drivers performed worst on question 8 ("What do the long white lines along the centre of the road mean?"), with 46 out of 60 experienced drivers not realising that the central lines act as a hazard warning.

Correlating official test scores with scores from the new tests

Only a small number of experienced drivers could remember their official theory test score or their official hazard perception test score (if they had done one). Accordingly, we only correlated learner test scores where these measures were concerned. Some learner drivers had not yet taken a theory or hazard test at the point of participating, and others could not remember their scores. This led to some missing cells and variable Ns across the correlations.

Only two correlations between test scores reached significance (Table 4). The correlation between the official theory test score and the official HP score is small, but suggests that learners who performed well in one test tended to perform well in the other. This perhaps reflects a general level of preparedness for the tests. More interestingly, the experimental theory questions correlate well with scores on the official theory test, while the correlation of HP score on the experimental test and the official HP test score only narrowly failed to reach significance (possibly due to the low number of learner drivers who were able to provide us with both measures). The hazard prediction test clearly does not correlate with the official hazard perception scores provided by participants.

Table 4. Pearson R correlation coefficients (with Ns and p-values) comparing experimental test scores to participants' self-report scores on the official tests. The between-group design did not allow for a correlation between our experimental perception and prediction scores. Correlations involving 'official' scores only included learner drivers. Significant correlations are in bold.

	Official Hazard Perception Test score	Experimental Theory Test score	Experimental Hazard Perception score	Experimental Hazard Prediction score	Driving-related video game play
Official Theory Test score	.328 **p = .034**	.565 **p < .001**	.377 p = .11	.250 p = .24	-.068 p = .69

	N = 42	N = 43	N = 19	N = 24	N = 37
Official Hazard Perception score		.120 *p* = .44 N = 43	.414 *p* = .07 N = 20	.043 *p* = .85 N = 23	-.452 ***p* = .005** N = 37
Experimental Theory Test score			-.07 *p* = .61 N = 58	.20 *p* = .12 N = 60	-.082 *p* = .40 N = 109
Experimental Hazard Perception score				N/A	-.310 ***p* = .026** N = 52
Experimental Hazard Prediction score					-.043 *p* = .75 N = 57

One other measure was included in the correlations: driving-related video game play. This was a rating reflecting participants' level of engagement with this form of entertainment, from 1 to 5. The current correlations, albeit on a highly restricted sample, suggest that the official HP test is negatively correlated with driving game play: The more frequently participants report playing driving games, the lower their score on the official HP test. Our CGI hazard perception test also correlated negatively with participants' game-play ratings. This may reflect a higher threshold for reporting hazards in players of driving games, compared to those participants who do not play such games. The hazard prediction test does not show a significant correlation with game play however.

Participants' thoughts on hazard perception tests

Following the study, participants were given an evaluation questionnaire containing a number of questions that asked their thoughts on the experimental tests. These questions were presented as semantic differentials on a 1 to 7 scale. This section details participant responses to these questions. Please note that the degrees of freedom vary across the following analyses, as some participants did not answer every question.

Do you prefer a hazard test to contain video-recorded clips or CGI-rendered clips?

This question asked whether participants prefer clips to be video-recorded or created in CGI, using a 7-point semantic differential scale. All participants were exposed to CGI content in the current study. Most of the participants also had experience of video-recorded hazards either through sitting the official UK test prior to 2015 (when the CGI version was introduced), through online training resources, or general awareness.

When analysed via a 2 x 2 ANOVA there was no effect of driver group or test-variant, nor an interaction, with an overall mean rating of 4. Though some participants had strong preferences one way or the other, the modal response was also around the middle of the scale (4 for learners, 5 for experienced drivers, tipping towards a preference for CGI).

Do you prefer a hazard test to contain many short clips or a single, long clip?

All hazard perception tests use clips of less than 60 seconds, containing one hazard (or occasionally two, in the official UK test). Most participants had undertaken a test using multiple short clips, or were at least aware that this is the format of the official test. Participants' ratings from 1 to 7 (with higher values favouring a single, long clip with multiple hazards) were subjected to a 2 x 2 between-groups ANOVA. This revealed a main effect of group, with experienced drivers favouring a single, long clip, while learners were more evenly distributed around the centre of the scale ($F_{(1,114)} = 21.4$, $MSe = 3.4$, $p < .001$; with mean ratings of 5.5 vs. 4.0). There was no effect of test-type, nor an interaction.

Do you prefer separate hazard/theory tests or a combined test?

This 7-point semantic differential assessed participants' preference for having separate tests of theory questions and hazard perception (as occurs in the official UK licensing procedure), or whether they preferred the combined variant used in the current study. The 2 x 2 ANOVA did not find any differences between the groups or test-variants. The mean rating was 5.3 and the modal response was 7, suggesting that most participants strongly favoured a combined test.

Was the overall experience better or worse than the official HP test?

The 2 x 2 ANOVA on this semantic differential found no effect of driver experience, nor an interaction, though it did reveal a main effect of test-variant, with participants rating the experimental hazard perception test higher (in comparison to the official test), than the experimental hazard prediction test ($F_{(1,73)} = 7.01$, $MSe = 1.60$, $p = .01$; with means of 5.4 and 4.5, respectively). Modal responses for the perception and prediction test was with modal responses of 6 and 4, suggesting that most participants thought the experimental hazard perception test was better than the official test, though they believed the hazard prediction test to be of a comparable experience to the official test.

Is our test more or less representative of real-life compared to the official HP test?

A 2 x 2 ANOVA did not find differences between the groups, or test variants, nor an interaction. The overall mean and modal responses suggest that participants thought that the experimental tests were more representative of real-world driving than the current official test (with a mean of 5.2 and a mode of 6).

Is our test more or less enjoyable than the official HP test?

A 2 x 2 ANOVA was conducted on this 7-point semantic differential. The analysis did not find any effects, nor an interaction, though the overall mean and modal responses suggest that participants thought that the experimental tests were more enjoyable than the official test (5.3 and 5, respectively).

Free-response comments

Following the semantic differentials, participants were provided with a space for free-response comments. Those who chose to write comments (who were mostly learners) were all positive about the new tests. These comments included: “Very good, good use of time and practice” (Learner, hazard perception test); “It was better… I preferred that it had a voice over” (Learner, hazard perception test); “It was more engaging, a lot better to spot hazards” (Learner, hazard perception test); “I felt the experimental [test] was far more interesting than the current one, felt it kept you on your toes and it seemed more realistic” (Learner, hazard prediction test); “Very different, it was more difficult than small clips but it was interesting to see how it works” (Learner, hazard prediction test); “The standard UK hazard perception test is less realistic compared to the experimental one. Although the experimental one is a bit more challenging, it is more like real life driving experiences.” (Learner, hazard prediction test).

These selected comments represent the wider corpus, with underlying themes of positivity towards the new tests, recognition of the benefit of the directional voice-over, and reports of heightened engagement and feelings of realism (even in the prediction task where the visual world is occluded at the point of hazard onset). For the hazard prediction test, several participants noted that this was a more difficult form of assessment, but this did not diminish their positivity.

Discussion

The results clearly demonstrate that both forms of test (hazard perception and hazard prediction) produce significant behavioural differences between our driver groups, with experienced drivers significantly outperforming the learners. Typically, effect sizes of 0.2, 0.5 and 0.8 are considered to reflect small, medium and large effects (Cohen, 1977). When comparing the total score on the hazard perception test across driver groups, this traditional scoring method produced an effect size of 1.19, which is almost unheard of in the hazard perception literature. Certainly, our previous research has never found such a large effect using response times, or any derivative of response times.

The hazard prediction test produced a respectable effect size of 0.7 (which is close to being a large effect). This supports the previous studies that have found the hazard prediction test format to differentiate successfully between driver groups. Our initial hypothesis was that the superiority of the prediction format noted in naturalistic video-based tests (Crundall and Kroll, 2018, Ventsislavova et al., 2019) would also be apparent with our new CGI clip. This directional hypothesis was not supported. Instead, the results follow those of Malone and Brünken (2016): while both tests differentiated between the driver groups, the effect size was larger for the traditional response time measure.

When the hazard perception test was rescored to reflect the percentage of hazards correctly identified (in order to compare it to the prediction test), it gave an effect size of 0.65, almost identical to that of the prediction test. A multilevel logistic regression, accounting for variance across participants and hazards, confirmed clear experiential effects for both tests, though the lack of interaction did not support the superiority of the prediction test. These results suggest that the prediction test is, at best, on a par with the perception test, and if one considers the effect size calculated from the 5-point scores per hazard, then the hazard perception test has the edge.

A number of differences between the current study and those of Ventsislavova et al., (2019) and Crundall and Kroll (2018) may explain why the current prediction test did not out-perform the perception test. The potential impact of using designed CGI animated hazards is the most obvious step-change in methodology, and it is worth considering what differential impact these animated hazards might have had on our two tests.

Are designed hazards suitable for hazard prediction?

In regard to the use of designed hazards, there is always the possibility that the resultant event contains only the coarsest of behaviours, and none of the subtle behavioural or environmental cues that might lead to a correct prediction. Even if the designers (in our case, traffic psychologists) have complete insight into all the relevant precursors for a particular hazard (which is by no means a certainty), they also need this knowledge to be declarative, so as to communicate this fully to the animators. The animators then interpret their instructions within the restrictions they have in terms of feasibility, time and budget. Thus, there are many stages in the development of high-fidelity CGI clips, in which errors of insight, communication and comprehension can contribute to a less than perfect hazard. It is possible to reduce the potential for errors by reducing the communication chain: Malone and Brünken (2015, 2016) presumably created their own CGI clips using VICOM software (www.vicomeditor.de), which is a powerful user-friendly animating package designed to generate driver-training materials. If the traffic researchers can create their own clips then they remove the potential for communication errors between designers and animators. Inevitably, however, if software is simple enough for non-animators to use with minimal training, it will not produce clips of the level of complexity that could be produced by professional animators. Thus, while we can reduce errors in the chain of communication, we potentially increase deviations from reality in the resultant clips due to necessary limits of user-friendly software and our non-professional skills.

For these reasons, naturally captured hazards on video could be argued to better reflect the subtleties of the precursors. Indeed, we do not need to rely on experts to prescribe the subtle precursors for inclusion in a video-based test, because those subtleties appear in the real world. For instance, if we observe the dip of the front end of another car when braking, this may reinforce our perception that the braking was harsh. This in turn may lead us to believe that the braking driver could be angry with the other road user that precipitated the emergency response. If we had achieved this level of subtlety in hazard 2 for instance, we may have seen higher prediction rates for the subsequent hazard (where the frustrated driver suddenly overtakes the vehicle that has caused him to brake sharply).

This example also shows why the loss of subtle cues would be more likely to impact a hazard prediction test than a hazard perception test. When the frustrated driver makes the dangerous overtaking manoeuvre in hazard 2, it is clear to the majority of participants in the hazard perception test that this is a hazard worth responding to (Figure 5, Hazard 2), even though the speed with which they respond is relatively slow (Figure 3, Hazard 2). In the hazard prediction test however, accuracy rates are much lower suggesting that both experienced and learner drivers had problems predicting what would happen next (with performance hovering around chance). It is possible that if this hazard had been captured on video, replete with all precursor subtleties, prediction performance would have been higher. More importantly, differentiation between the groups may have also been higher, as the experienced drivers may be more sensitive to these subtleties than the learners.

There are however valid counter-arguments in favour of CGI clips. First, one might consider it churlish to complain about CGI given the substantial experiential effect sizes found in this study for both the hazard perception and hazard prediction tests. Second, naturalistic hazards rarely look as good in reality as in theory. Sun glare, rain, poor camera angles, insufficient distraction, too much distraction, and any hazard mitigation on the part of the film-car driver can all diminish the quality of a clip. Animation can remove all of these problems and present the hazard in an approximation of the best possible conditions. A further benefit is that CGI animations can be iterated following behavioural testing to improve their ability to differentiate between safe and less-safe drivers, though in reality, budget and time constraints often prevent this from happening.

These issues with designed hazards for hazard prediction tests are not insurmountable. Continued improvement in both animation techniques and in our understanding of hazardous precursors will improve our ability to assess hazard prediction skill, and build on the significant results found in this study. The use of high-definition video-based tests of hazard prediction skill also remains as an attractive option.

A further design issue for the current study was the ordering and temporal proximity of the events. To prevent participants detecting a pattern in the presentation of hazards and events, we avoided a strict ‘Hazard > Question > Hazard > Question’ ordering of events, allowing successive questions and hazards to disrupt this ordinal pattern. Consideration was also given to gap between events. Hazards and questions could not appear at regular temporal intervals without giving participants cues to the timing of forthcoming events. Thus, some events were designed to appear closer together in time than others. The impact of a preceding event upon the detection of a subsequent hazard could be a concern however. To avoid this, any hazard was not preceded by an event for a minimum of 15 seconds, which is easily sufficient to overcome carryover effects such as the attentional blink (Petersen, Kyllingsbæk, & Bundesen, 2012) or the impact of a preceding target on subsequent search strategies (Thompon, Howting and Hills, 2015). It remains possible however that rumination on a preceding hazard or theory question could have overlapped with the onset of the next hazard.

The temporal gap between a hazard and a subsequent question was much shorter in some cases. For instance, the gap between the onset of Hazard 4 and Question 4 is less than six seconds. Such short gaps were considered appropriate where the question related directly to the hazard (in this case, the theory question concerned the flashing of one’s headlights; an action which precipitated the preceding hazard). While 26 out of 60 experienced drivers failed to answer this question correctly, it is unlikely that unrelated carry-over effects are the cause of this poor performance. It is more likely that many experienced drivers hold the incorrect view of when headlights should be flashed (i.e. as a warning, rather than to give way to other road users), despite the clear indication from the preceding hazard that this view is dangerous.

Are hazard perception and hazard prediction tests measuring different things?

Beyond the adoption of situation awareness, there is a lack of developed theory underlying hazard perception (Horswill, 2017). Nonetheless, it is tempting to think of hazard prediction as a measure that comes earlier in the chain of behaviours that lead to hazard avoidance, and thus avoids any post-perceptual confounds (Pradhan and Crundall, 2017). On this basis, active prediction of hazards should lead to faster perception of hazards when they occur, resulting in faster response times in a traditional test. This would imply that those hazards that are responded to most frequently in the hazard perception test, should also be the hazards that are most easily predicted, and vice versa. Figure 5, however, suggests this not to be the case in at least two of the hazards where group differences in one test were not reflected in group differences in the other test.

Accepting that this was a between-groups design, might this suggest that the prediction and perception tests are measuring different aspects of hazard avoidance skill? For instance, in hazard 5 there was no evidence of a group difference on the prediction task, yet a performance gap was apparent between groups in the perception task. Does this mean that prediction was not necessary for drivers to perceive the hazard in the perception test? A look at the incorrect options that drivers chose in the prediction task suggests this was not the case. The correct answer (see Table 2) was that a pedestrian with a buggy would emerge from between parked cars on the left. Only 30% of all participants in the prediction test chose this option. A further 45% however chose the distracter option: ‘The white parked car on the left tries to pull off as you pass it’. These two options, accounting for 75% of responses, refer to the same spatial location within the scene. Thus, it appears that the majority of the participants had successfully predicted the location of the imminent hazard, even though they had not gleaned enough information to allow precise identification. A similar process presumably encouraged participants in the hazard perception test to apply a *scene prior* to the location of the parked cars, in anticipation of an unspecified hazard (Toralba et al., 2007), resulting in a greater likelihood of a response within the time window once the hazard became visible. Instead of arguing for prediction as a distinct process to perception, this finding suggests that prediction works at multiple levels, and that predicting the location of a potential hazard can still be beneficial even if the exact nature of the hazard is ambiguous.

A further argument for prediction and perception measuring separate underlying processes could be interpreted in terms of the correlations. While our experimental perception test showed a marginal relationship with participants' scores on the official test, there was clearly no relationship between official test scores and participants' hazard prediction score. Yet again however, this is more likely due to the different way of measuring hazard skill than differences in the underlying measure.

Regardless of the reason for the underlying differences across hazards, the results suggest that, in their current state, some hazards would be better suited for eliciting response-time measures, while other hazards may fare better in a prediction test (in terms of their ability to differentiate between driver groups). For instance, as noted above, multiple precursors may prime the same spatial location leading to poor prediction scores but improving response times to a subsequent hazard. Alternatively, multiple potential hazards that overlap a scoring window may raise the possibility that a simple button response may be scored as a 'hit', though in reality it was made in response to a precursor that did not develop into a hazard (i.e. in truth, a 'miss'). This raises the interesting possibility that a future test might benefit from a hybrid approach.

The inclusion of theory questions

Performance on the theory questions was not impacted by the test variant. Neither was there a difference between the driver groups. This perhaps reflects the preparation of learner drivers for the driving test offsetting any experiential benefit in the experienced driver group. The fact that performance on hazard detection did differ between groups, but theory test performance did not, underlines the difficulty of imparting hazard perception skills to learner drivers. While theory questions test declarative knowledge for a range of facts and rules that are relatively easy to learn. Hazard perception, however, is a skill that cannot yet be reduced to an instruction manual.

In addition to the rehearsed knowledge of learner drivers, the failure to find a group difference in the theory questions may also reflect a degradation of knowledge in experienced drivers over time, or the possible overwriting of knowledge due to real-world exposure. Particularly worrying is the 40% of experienced drivers who incorrectly believed that flashing headlights indicate that another road user is giving way to you. While this occurs frequently on UK roads, the UK Highway Code states that flashing one's headlights should only be used as a warning. It is possible that this incorrect response was primed by the actions of an oncoming driver in the CGI clip who did flash their headlights (which ultimately led to hazard 4), supporting the use of context to elicit more realistic responses.

The strong correlation between learners' scores on our theory questions, and their scores on the official test suggests that our combination of a theory test with hazards was not detrimental to their performance. Neither was it detrimental to the ability of the hazard tests to differentiate between driver groups. Indeed, while practicalities rendered it impossible to compare an isolated hazard test to a combined hazard/theory test in the current study, the healthy effect sizes for both hazard tests may in part be due to the additional demands placed on drivers through the inclusion of the theory questions. Furthermore, the modal response of participants giving a preference for a combined hazard/theory test compared to two separate tests was at the highest point of the semantic differential (7 on a scale from 1 to 7), suggesting that all participants were extremely positive about this new combined test format.

The influence of video-game playing

As previously noted, studies have reported positive relationships between video game play and a range of perceptual, cognitive and motor skills, including driving (e.g. Rupp et al., 2015), though others have found a negative relationship between video game play and risk-taking in driving scenarios (e.g. Deng et al., 2017). In regard to hazard perception tests, Vlakveld (2014) found video game play positively correlated with hazard performance. In contrast, our results suggest that video game pay negatively correlates with both performance on the official UK hazard perception test and our experimental hazard perception test. It should be noted however that our measure of video game play was specific to driving-related games, whereas Vlakveld's question was not. This may partially account for the difference in findings across the two studies, though cross-cultural differences in how one relates to video games may also play a difference (e.g. Anderson et al., 2007).

Why might driving-specific video game play reduce hazard perception performance, especially when video game play is supposed to improve vigilance for infrequent targets (Szalma et al., 2018)? One suggestion is that video-game play can desensitize players to the threat level of stimuli. For instance, violent video games can impair threat perception based on others' facial emotions (Denson, et al., 2020), and can diminish the typical attentional blink found after viewing threatening or otherwise aversive stimuli (Jin et al., 2018). Most driving games rely on speed and outlandish danger to impart fun and enjoyment. It is possible that these elements desensitise game players to real-world driving hazards, perhaps impacting on their criterion bias (i.e. raising their threshold) for determining something to be hazardous. The failure to find a correlation between video-game play and our hazard prediction test fits with this post-hoc rationalisation, as the prediction test was initially designed to mitigate the possible confound of criterion bias in more traditional hazard perception tests (Crundall, 2016). This does not mean that video-game players' higher threshold for reporting hazards is not important, but such problems with appraisal should be measured distinctly from the ability to spot the hazard in the first place. The lack of relationship between hazard prediction and game playing also suggests that the use of CGI clips is not responsible for the relationship between game playing and hazard perception scores. If this were the case, the relationship should also have been noted in the hazard prediction scores. Further research is required to identify what key experiences from driving-related video game play evoke this negative relationship.

Participant preferences

All of our innovations were received with varying levels of approval from participants. None of the mean or modal responses regarding either the hazard perception or hazard prediction test dropped below the mid-point of the semantic differential scales. Both tests were considered more realistic and enjoyable than the current official UK test, though the hazard perception test was reported to give the better overall experience. This may have reflected the more disjointed experience of the prediction test, with hazard occlusions disrupting the flow of the experience (although theory questions also disrupted the flow in both tests). Alternatively, it may have reflected the increased difficulty of the hazard prediction test, which may have led drivers to either question their own skills, or (perhaps defensively) question the veracity of the test. Future research should look to unpack this ambivalence towards the prediction test used in this study. Regardless of the differences between ratings of the hazard perception and the hazard prediction test, the latter test was rated as at least providing a comparable experience to the official test.

Of all the other innovations, the most popular was the combination of hazards and theory questions into a single test. The use of the long clip (compared to traditional short clips) and the use of CGI (compared to the historical use of video) produced more ambivalent ratings, though means and modes never dipped below the middle of the scale. Free responses were all in favour of the new tests compared to the current official tests, even whilst acknowledging hat the prediction test was harder than they expected. These ratings and responses provide a snapshot of public reaction might be if changes were made to the national test along these lines. The combination of theory and hazard tests would likely be met with the least public resistance.

Conclusion

This study is the first to compare a hazard perception test to a hazard prediction test using high-fidelity CGI animations, comparable to those used in the official UK test. It follows the methodology of Crundall and Kroll (2018) and Ventsislavova et al. (2019), but does not replicate their results: Instead of declaring the hazard prediction test to be the winner, the current study found both tests to perform well in differentiating between driver groups. Indeed, the effect size of the hazard perception test was exceptionally large, suggesting that the CGI animations are particularly suited to tests involving speeded responses to hazards (similar to Malone and Brünken, 2016).

Despite the comparable performance of the two test-variants, there are other compelling reasons to use a prediction format, such as the opportunity to have drivers consider multiple hazards for a particular scenario. This may avoid learners becoming preoccupied with one potential outcome that could influence their on-road safety. For instance, if they always associate a parked bus with a hidden pedestrian, they may be caught by surprise on the occasion that a pedestrian crosses from the opposite side of the road in order to catch the bus. Furthermore, assessment predicates training. If we merely assess the ability to respond to hazards, drivers will train towards improving their response times, with strategies that may not develop the underlying real-world skill. However, assessment on prediction skills, will encourage drivers to seek specific training in hazard prediction. As prediction likely underlies real world hazard avoidance, this should have greater benefit. Finally, we must note that scores on the experimental hazard perception test negatively correlated with self-report play on driving video games, whereas hazard prediction scores did not. While these correlations need replication, any relationship with video game play should raise concerns that the test is tapping into something other than driving safety skills.

Of the other innovations in this study, the combination of the theory questions with hazards appealed most to our participants. Furthermore, the correlation between the experimental theory questions and their official theory scores (where participants could remember them), suggests that there is unlikely to be a negative impact on typical theory question performance from combining these tests. Indeed there are logical arguments that providing questions embedded in context will provide a more realistic assessment of whether drivers have access to this information at the time when it is most needed.

One interesting finding of this study however is that some of our hazards better suited a prediction test, while others were better served by a perception test. Participants' choice of distracter answers in the prediction test (e.g. Hazard 5) give clear guidance to potentially improve the test, though it is also likely that CGI animations need to include more subtle precursors to fully embrace the potential of the prediction format. Alternatively, we may choose to embrace the differences in the current hazard set, and seek a hybrid model that may capture the best of both methodologies.

References

Achtman, R. L., Green, C. S., & Bavelier, D. (2008). Video games as a tool to train visual skills. *Restorative neurology and neuroscience*, *26*(4, 5), 435-446.

Adminaite, D. J., Stipdonk, G., & H & Ward, H. (2017). Ranking EU progress on road safety: 11th road safety Performance Index (PIN) report.

Al-Aamri, A. K., Padmadas, S. S., Zhang, L. C., & Al-Maniri, A. A. (2017). Disentangling age–gender interactions associated with risks of fatal and non-fatal road traffic injuries in the Sultanate of Oman. *BMJ global health*, *2*(3).

Anderson, C. A., Gentile, D. A., & Buckley, K. E. (2007). *Violent video game effects on children and adolescents: Theory, research, and public policy*. Oxford University Press.

Borowsky, A., Shinar, D., & Oron-Gilad, T. (2010). Age, skill, and hazard perception in driving. *Accident Analysis & Prevention*, *42*(4), 1240-1249.

Boufous, S., Ivers, R., Senserrick, T., & Stevenson, M. (2011). Attempts at the practical on-road driving test and the hazard perception test and the risk of traffic crashes in young drivers. *Traffic injury prevention*, *12*(5), 475-482.

Castro, C., Padilla, J. L., Roca, J., Benítez, I., García-Fernández, P., Estévez, B., ... & Crundall, D. (2014). Development and validation of the Spanish hazard perception test. *Traffic injury prevention*, *15*(8), 817-826.

Ciceri, M. R., & Ruscio, D. (2014). Does driving experience in video games count? Hazard anticipation and visual exploration of male gamers as function of driving experience. *Transportation research part F: traffic psychology and behaviour*, *22*, 76-85.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic press.

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

Deng, M., Chan, A. H., Wu, F., & Liu, S. (2017). Effects of the contextual variables of racing games on risky driving behavior. *Games for health journal*, *6*(4), 249-254.

Denson, T. F., Dixson, B. J., Tibubos, A. N., Zhang, E., Harmon-Jones, E., & Kasumovic, M. M. (2020). Violent video game play, gender, and trait aggression influence subjective fighting ability, perceptions of men's toughness, and anger facial recognition. *Computers in Human Behavior*, *104*, 106175.

Dye, M. W., Green, C. S., & Bavelier, D. (2009). The development of attention skills in action video game players. *Neuropsychologia*, *47*(8-9), 1780-1789.

Elvik, R. (2010). Why some road safety problems are more difficult to solve than others. *Accident Analysis & Prevention*, *42*(4), 1089-1096.

Endsley, M. R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. *Situation awareness analysis and measurement*, *10*, 147-173.

Gartenberg, D, Gunzelmann, G., Hassanzadeh-Behbaha, S., Trafton J. G., (2018). Examining the role of task requirements in the magnitude of the vigilance decrement. *Frontiers in Psychology, 9*, 1054.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534-537.

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological science*, *18*(1), 88-94.

Horswill, M. S., Marrington, S. A., McCullough, C. M., Wood, J., Pachana, N. A., McWilliam, J., & Raikos, M. K. (2008). The hazard perception ability of older drivers. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *63, 4*, 212-218.

Horswill, M. S., Hill, A., & Wetton, M. (2015). Can a video-based hazard perception test used for driver licensing predict crash involvement?. *Accident Analysis & Prevention*, *82*, 213-219.

Horswill, M. S. (2016). Hazard perception in driving. *Current Directions in Psychological Science*, *25*, 6, 425-430.

Horswill, M. S. (2016). Hazard perception tests. In Donald L. Fisher, Jeff K. Caird, William J. Horrey and Lana M. Trick (Ed.), *Handbook of teen and novice drivers: Research, practice, policy, and directions* (pp. 61-74). Boca Raton, Florida, United States: CRC Press. doi:10.1201/9781315374123

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., & McKenna, F. P. (2004). Drivers' hazard perception ability: Situation awareness on the road. In S. Banbury & S. Tremblay (Eds.). A Cognitive Approach to Situation Awareness (pp.155-175). Aldershot, UK: Ashgate.

Howard, C., Wilding, R., & Guest, D., (2016). Light video game play is associated with enhanced visual processing of rapid serial visual presentation targets. *Perception, 46, 2,* 161-177.

Huestegge, L., Skottke, E. M., Anders, S., Müsseler, J., & Debus, G. (2010). The development of hazard perception: Dissociation of visual orientation and hazard processing. *Transportation research part F: traffic psychology and behaviour*, *13, 1*, 1-8.

Insurance Institute for Highway Safety (IIHS), (2019). Fatality Facts: Teenagers. Retrieved from https://www.iihs.org/iihs/topics/t/teenagers/fatalityfacts/teenagers (last accessed on 16/07/20).

McDonald, C. C., Goodwin, A. H., Pradhan, A. K., Romoser, M. R. E., & Williams, A. F. (2015). A review of hazard anticipation training programs for young drivers. *Journal of Adolescent Health, 57, 1, (Suppl),* S15–S23.

McKnight, A.J., & McKnight, A.S. (2003). Young novice drivers: Careless or clueless? *Accident Analysis & Prevention*, *35*(6), 921–925.

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.

Ismail, R., Voon, N. L., Saad, M. H. M., Saleem, M., & Ibrahim, N. (2016). Aberrant driving among young malaysian drivers. *Jurnal Teknologi*, *78*(6-10).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.

Jin, M., Onie, S., Curby, K. M., & Most, S. B. (2018). Aversive images cause less perceptual interference among violent video game players: evidence from emotion-induced blindness. *Visual Cognition*, *26, 10,* 753-763.

Kinear, N., Lloyd, L., Helman, S., Husband, P., Scoons, J., Jones, S., Stradling, S., McKenna, F., & Broughton, J. (2013). *Novice drivers – evidence review and evaluation*. TRL Report, PPR673. Transport Research Laboratory, Crowthorne

Kroll, V., Mackenzie, A. K., Goodge, T., Hill, R., Davies, R., & Crundall, D. (2020). Creating a hazard-based training and assessment tool for emergency response drivers. *Accident Analysis & Prevention*, *144*, 105607.

Malone, S., & Brünken, R. (2015). Hazard perception assessment–How much ecological validity is necessary? *Procedia Manufacturing*, *3*, 2769-2776.

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.

McGowan, A., & Banbury, S. (2004, September). Interruption and reorientation effects of a situation awareness probe on driving hazard anticipation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, No. 3, pp. 290-294). Sage CA: Los Angeles, CA: SAGE Publications.

McKenna, F. P., & Crick, J. L. (1994). *Hazard perception in drivers: A methodology for testing and training*. TRL Report CR313. Transport Research Laboratory, Crowthorne.

Moafian, G., Aghabeigi, M. R., Hoseinzadeh, A., Lankarani, K. B., Sarikhani, Y., & Heydari, S. T. (2013). An epidemiologic survey of road traffic accidents in Iran: analysis of driver-related factors. *Chinese journal of traumatology*, *16, 3*, 140-144.

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.

Pattyn N., Neyt X., Henderickx D., & Soetens E. (2008). Psychophysiological investigation of vigilance decrement: Boredom or cognitive fatigue? Physiology and Behavior, 93, 369–378.

Petersen, A., Kyllingsbæk, S. & Bundesen, C., (2012). Measuring and modeling attentional dwell time. *Psychonomic Bulletin and Review, 19,* 1029–1046.

Pradhan, A. K., & Crundall, D. (2016). Hazard avoidance in young novice drivers: Definitions and a framework. In Donald L. Fisher, Jeff K. Caird, William J. Horrey and Lana M. Trick (Ed.), *Handbook of teen and novice drivers: Research, practice, policy, and directions* (pp. 439-450) Boca Raton, Florida, United States: CRC Press. doi:10.1201/9781315374123

Rupp, M. A., McConnell, D. S., & Smither, J. A. (2016). Examining the relationship between action video game experience and performance in a distracted driving task. *Current Psychology*, *35, 4*, 527-539.

Scialfa, C. T., Borkenhagen, D., Lyon, J., Deschênes, M., Horswill, M., & Wetton, M. (2012). The effects of driving experience on responses to a static hazard perception test. *Accident Analysis & Prevention*, *45*, 547-553.

Szalma, J. L., Daly, T. N., Teo, G. W. L., Hancock, G. M., & Hancock, P. A. (2018). Training for vigilance on the move: A video game-based paradigm for sustained attention. *Ergonomics*, *61, 4*, 482-505.

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* (No. DOT HS 812 235). Dunlap and Associates, Inc..

Thompson, C., Howting, L., & Hill, P., (2015). The transference of visual search between two unrelated tasks: Measuring the temporal characteristics of carry-over. *The Quarterly Journal of Experimental Psychology, 68, 11,* 2255-2273.

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. H., (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review, 113, 4,* 766-786.

Underwood, G. (2007). Visual attention and the transition from novice to advanced driver. *Ergonomics*, *50*(8), 1235-1249.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., ... & Li, Q. (2019). A comparison of hazard perception and hazard prediction tests across China, Spain and the UK. *Accident Analysis & Prevention*, *122*, 268-286.

Vlakveld, W. P. (2014). A comparative study of two desktop hazard perception tasks suitable for mass testing in which scores are not based on response latencies. *Transportation research part F: traffic psychology and behaviour*, *22*, 218-231.

Vogel, K., Kircher, A., Alm, H., & Nilsson, L. (2003). Traffic sense—which factors influence the skill to predict the development of traffic scenes?. *Accident Analysis & Prevention*, *35, 5*, 749-762.

Wayne, N. L., & Miller, G. A. (2018). Impact of gender, organized athletics, and video gaming on driving skills in novice drivers. *PloS one*, *13, 1*, e0190885.

Wells, P., Tong, S., Sexton, B., Grayson, G., & Jones, E. (2008). *Cohort II: a study of learner and new drivers. Volume 1—Main Report* (Vol. 81). Road Safety Research Report.

Footnotes

[1] Some questions include a static picture to provide rudimentary context. The UK Driver and Vehicle Standards Agency are planning to add some small video clips to a selection of questions in the official test in the near future.

[2] These means are shrinkage estimates from a multilevel model and can be interpreted as estimates from a typical person on a typical question. In effect, they partially pool information across items and individuals give less weight to atypical questions or individuals.