Running Header: HAZARDS AND PROFESSIONAL DRIVERS

Prediction and perception of hazards in professional drivers: Does hazard perception skill differ between safe and less-safe fire-appliance drivers?

David Crundall and Victoria Kroll

Nottingham Trent University

Author Note

David Crundall and Victoria Kroll, Department of Psychology, School of Social Sciences, Nottingham Trent University, UK.

This work was supported by The Fire Service Research and Training Trust, UK [2015-16]. We would like thank Craig Parkin, Paul Foster, Ivan Hayes, Vivienne Brunsden and Mark Jarman for all of their assistance, and all of the emergency response drivers (and control participants) who gave up their time to participate.

Correspondence concerning this article should be directed to David Crundall (david.crundall@ntu.ac.uk).

Abstract

Can hazard perception testing be useful for the emergency services? Previous research has found emergency response drivers' (ERDs) to perform better than controls, however these studies used clips of normal driving. In contrast, the current study filmed footage from a fire-appliance on blue-light training runs through Nottinghamshire, and endeavoured to discriminate between different groups of EDRs based on experience and collision risk. Thirty clips were selected to create two variants of the hazard perception test: a traditional push-button test requiring speeded-responses to hazards, and a prediction test that occludes at hazard onset and provides four possible outcomes for participants to choose between. Three groups of fire-appliance drivers (novices, low-risk experienced and high-risk experienced), and age-matched controls undertook both tests. The hazard perception test only discriminated between controls and all FA drivers, whereas the hazard prediction test was more sensitive, discriminating between high and low-risk experienced fire appliance drivers. Eye movement analyses suggest that the low-risk drivers were better at prioritising the hazardous precursors, leading to better predictive accuracy. These results pave the way for future assessment and training tools to supplement emergency response driver training, while supporting the growing literature that identifies hazard prediction as a more robust measure of driver safety than traditional hazard perception tests.

Keywords: hazard perception; hazard prediction; professional drivers; fire service; fire appliance drivers; emergency response driving.

Introduction

A Brief Overview of Hazard Perception

Hazard perception (HP) skill is the ability of a driver to detect on-road hazards that could cause a potential collision, and it is claimed to be the only higher-order cognitive skill that reliably relates to crash risk in drivers (Horswill and McKenna, 2004). This skill is typically measured using video clips of real driving filmed from the driver's perspective, from a windscreen or roof-mounted video camera. The driver watches the video clips on a computer and must make a response (usually a simple button press) to any perceived hazard. The speed of the button press is the typical primary measure of judging driver safety, based on the simple premise that if drivers can spot on-road hazards quickly, they are more likely to avoid them. There have been a number of studies that have found hazard perception tests to discriminate between experienced, safer drivers and novice, or less-safe, drivers (e.g. Pelz & Krupat, 1974; Watts & Quimby, 1979; McKenna and Crick, 1991; Deery, 1999; Wallis & Horswill, 2007; Horswill et al., 2008; Pradhan et al., 2009; Horswill, Taylor, Newnam, Wetton, & Hill, 2013; Scialfa et al., 2011). Performance on a hazard perception test has even been found to predict the likelihood of being involved in a future traffic collision (Drummond, 2000; Boufous et al., 2011), which supports suggestions that under-developed hazard perception skill contributes to the over-representation of novice drivers in the collision statistics (Horswill and McKenna, 2004; Maycock et al., 1991; Underwood, 2007).

While certain aspects of hazard perception testing have been questioned in the academic literature (e.g. Crundall et al., 2012), the UK Government found the evidence sufficiently compelling to bring in such a test as part of the driver licensing procedure in 2002. Six years later a Government-sponsored research team reported that the introduction of the hazard perception test had resulted in a significant decrease in the number of certain types of collision on UK roads (Wells et al., 2008). This was considered to be due to keeping exceptionally poor drivers off the roads, while encouraging the average learner driver to practice the higher-order cognitive tasks involved in predicting and responding to on-road hazards.

Hazard perception in the emergency services

If experience and training lead to improved hazard perception performance, one might imagine that those professional drivers, who are trained to drive under extreme conditions, such as emergency response drivers, should display the greatest levels of hazard perception ability. Indeed, several studies have compared ambulance drivers and police drivers to control groups, and found that these professional drivers exhibit superior response times to hazards in video clips of everyday driving (Johnston & Scialfa, 2016; McKenna & Crick, 1991; Horswill et al., 2013). This superiority may reflect the fact that they are exposed to, and trained under, extreme conditions. Thus, when presented with a hazard perception test of normal driving clips, they find it relatively easy to identify the hazards, as the filmed driving occurs at a slower speed and involves more predictable manoeuvres than the emergency response scenarios they are regularly exposed to (see ‘above real time training’ for an approach that seeks to exploit this effect, Lorains, Ball, & MacMahon, 2013).

While these studies support the hypothesis that increased training and exposure can positively develop HP skill in normal driving conditions (though we acknowledge that self-selection may still play a part), they tell us nothing about how emergency service drivers cope with hazards in the line of duty. Travelling at speed relative to other traffic, contravening road rules, and influencing the actions of other road users via sirens and lights, are all likely to create hazards that the average driver will never need to worry about. A hazard perception test cannot assess emergency drivers’ abilities in detecting these hazards without using footage captured from realistic blue-light scenarios (i.e. filmed from a vehicle travelling under blue-lights and sirens). To the authors' knowledge, only one previous study has been published that used blue-light video footage, filmed from police cars involved in pursuit and emergency response situations (Crundall et al., 2003, 2005), which demonstrated that police drivers' eye movements and electrodermal responses differed to those of control drivers.

What all the above studies lack, however, is the opportunity to discriminate between safe and less-safe drivers within the emergency services. If HP skill is a cause of novice driver collisions, as put forward in the literature, then does this transfer to other end of the spectrum of experience (i.e. can HP skill still explain why some highly experienced drivers have collisions and others do not)? The findings of Horswill et al., (2013) certainly suggest that this could be the case. They demonstrated that even highly experienced drivers could benefit from hazard perception training, suggesting that HP skill might be a valuable diagnostic and training tool even within a group of professional emergency service drivers.

While the diagnostic efficacy of hazard perception tests at the upper end of the experience spectrum is an important theoretical question, it is also essential for the practical application of an HP test for the emergency services. The emergency services are not interested in demonstrating that their drivers are better than non-emergency service drivers at spotting hazards. They are, however, interested in identifying those emergency response drivers who are at risk, and could therefore benefit from additional training. Thus, a truly effective HP test should differentiate between emergency response drivers at different levels of risk, as well as experience, specific to their particular role. This is the aim of the current study: we want to assess whether HP skill can differentiate between professional driver groups, and design a test to capture this information for a specific sector of the emergency services: fire-appliance[1] drivers. This will expand our understanding of hazard perception as a skill that may or may not reach a plateau (Horswill et al., 2013), while simultaneously developing an HP test that can be used as a cost-effective supplement to on-road training and assessment in a service that faces high levels of risk on the roads (e.g. Becker et al., 2003; Crundall et al., 2003; Maguire et al., 2002) and stringent budget cuts in the UK (Chief Fire Officers Association, 2015).

Hazard perception or hazard prediction?

Pradhan and Crundall (2017) selected the term 'hazard avoidance' to describe the whole process of safely navigating a hazard. This includes a variety of sub-processes from searching for hazardous precursors and prioritising them for subsequent monitoring, through to processing, appraising, mitigating and responding to hazards when they occur. Hazard perception reflects a selection of these sub-processes, from visual search through to deciding whether the hazard really poses a threat. Unfortunately, this means that simple response times to an HP test confound several sub-processes. For instance, a hazard response does not just reflect how quickly one spots the hazard, but also how quickly one processed it, and, crucially, whether the hazardousness of the event reached an individual's threshold for reporting. The problem of criterion bias is especially concerning, as the most experienced drivers are likely to have a higher threshold for what constitutes a hazard. Thus while they may spot the hazard sooner than less-experienced drivers, they may wait to respond until the level of hazardousness has reached a relatively high threshold (Crundall, 2016). While we have briefly reviewed much research that has demonstrated the diagnostic abilities of hazard perception tests, there are also many studies that have failed to discriminate between driver groups with a simple push button response (e.g. Chapman and Underwood, 1998; Sagberg and Bjørnskau, 2006; Borowsky et al., 2010; Underwood et al., 2013). It is possible that criterion bias in experienced drivers may have caused these mixed findings.

As an alternative to a push-button response, we can directly measure when drivers spot hazards using eye tracking technology (and we have done so in the current study), but eye tracking is unsuitable for an assessment method intended for wide use. Instead, we may consider changing the nature of the test to isolate the key component of hazard perception skill. This has been the aim of a collection of studies that have developed an HP-variant called the ‘hazard prediction’ test. Based on the Situation Awareness Global Assessment Technique (SAGAT), the hazard prediction test presents drivers with a series of hazard clips that are suddenly occluded, just as the hazard begins to develop (Jackson et al., 2009; Castro et al., 2014; Crundall, 2016; Ventsislavova et al., 2016). Following occlusion, drivers are simply asked ‘what happens next?’. This test targets the driver’s ability to identify potential hazard precursors, and extrapolate the likelihood of them leading to a hazard (e.g. a high-sided lorry might hide a small child; a pedestrian walking along the sidewalk and glancing into the roadway, might step into the road, etc.). These precursors must be hierarchically prioritised and monitored accordingly, which will give the driver the best opportunity for identifying which one will actually develop into a hazard. Jackson et al., (2009) argued that the act of prediction is perhaps the most crucial aspect of hazard perception, as it primes both the location of future hazards and the ability to process them (though we acknowledge that the post-prediction processes also have a role to play).

One advantage of this approach is that it removes the need for drivers to compare an unfolding hazard to an internal criterion, which may then mask their ability to detect hazards compared to less-safe drivers. Instead of a confounded response time, we record the percentage accuracy of hazards successfully predicted. While the number of studies employing this HP-variant are still limited, the evidence suggests that this test is a robust discriminator of safe and less-safe drivers (Jackson et al., 2009; Castro et al., 2014; Crundall, 2016; Ventsislavova et al., 2016).

The first direct comparison of a hazard perception test with a hazard prediction test was recently undertaken across three countries: China, Spain and the UK (Ventsislavova et al., submitted)[2]. Novice and experienced drivers did not differ on the hazard perception test, but the test was found to be sensitive to the nationality of the participants, with Chinese drivers responding to fewer hazards than UK drivers. We suggested that this might reflect the higher hazard threshold of Chinese drivers who are typically exposed to a more hazardous driving environment. The hazard prediction test, however, provided the opposite results. Cultural differences between participants were reduced, while experienced drivers were found to out-perform novice drivers regardless of nationality. The results demonstrated that the hazard prediction test, when unconfounded by criterion level, appears to be a more robust and culturally-agnostic measure of driver safety.

Based on these data, one might be tempted to argue that the emergency services would be better served by a hazard prediction test rather than a hazard perception test. However, given the relative novelty of the hazard prediction test compared to the accepted success of the hazard perception test, we opted to create both a hazard perception test (experiment 1) and a hazard prediction test (experiment 2), in order to identify which is most suitable for discriminating between fire-appliance driver groups.

The current study

Multiple cameras were placed on a fire appliance (FA) to record footage of blue-light training runs through the city of Nottingham and the surrounding areas. From over 12 hours of footage, 30 clips were selected to create a hazard perception test and a hazard prediction test. The former required speeded responses to hazards (selected *a priori* from the footage), while the latter test required participants to identify ‘what happens next?’ by selecting one of four options following occlusion. Four groups of drivers were recruited to undertake both tests: A control group of non-emergency service drivers was used as a baseline, while three groups of FA drivers were defined as novices, high-risk experienced drivers and low-risk experienced drivers (based on the number, severity and blameworthiness of self-reported incidents). Comparisons of these different groups reflect different hypotheses. First, a comparison of control drivers to all FA drivers reflects the hypothesis that the advanced training and experience of all FA drivers should result in overall superior performance compared to average drivers, as noted in the literature (Johnston & Scialfa, 2016; McKenna & Crick, 1991; Horswill et al., 2013). Secondly, a comparison of novice FA drivers to the two groups of experienced FA drivers should reveal whether a basic experiential effect could be found. Given that even the 'novice' group would be still be considered as highly-experienced drivers under everyday conditions, this assesses whether experiential differences in HP tests are task (and therefore hazard) specific. Finally, the high and low-risk groups of experienced FA drivers were directly compared to assess whether their level of collision-involvement could be differentiated by the tests.

Experiment 1

The first experiment reports data from a traditional hazard-perception methodology. Four groups of participants (controls, novice FA drivers, experienced, high-risk FA drivers, and experienced, low-risk FA drivers) viewed a series of clips recorded from a fire appliance on a blue-light run, each containing one *a priori* hazard with a defined temporal scoring window. Participants had to press a button as soon as they saw a hazard. We predicted that control drivers would be slower than all FA drivers, that novice FA drivers would be slower than all experienced FA drivers, and that high-risk, experienced FA drivers would be slower than low-risk, experienced FA drivers. We also measured participants' eye movements with the hypothesis that these data would help explain any behavioural differences between the groups.

Method

Participants

Eighty-four drivers were assigned across four groups. The first group consisted of 21 novice fire-appliance drivers (18 male, 3 female) with a mean age of 35.4 years, 9571 personal miles per annum, and a mean personal driving experience of 16.5 years since passing their driving test. Owing to this being a challenging sample of participants to obtain, novice drivers were defined as fire fighters who were either currently completing the Emergency Fire-Appliance Driver (EFAD) course, or who were awaiting their EFAD course.

Forty-three participants were classed as experienced fire-appliance drivers (41 male, 2 female), with a mean age of 42.4 years of ages, a mean of 10.4 years' experience of fire appliance driving, a mean of 11069 personal miles per annum, and a mean driving experience of 23.4 years since passing the driving test. This sample was divided into high and low-risk groups on the basis of self-reported frequency, severity and blameworthiness of all recalled collisions across their driving history (including personal and at-work collisions). Severity ratings for each collision varied between 1 and 3 points, with 1 point reflecting a collision producing damage of less than £200 value, 2 points reflecting a collision producing damage of greater than £200 value, and 3 points for a collision resulting in an injury. Blame ratings also varied between 1 and 3 points, with 1 point reflecting the attribution 'not my fault', 2 points for 'partly my fault', and 3 points for 'completely my fault'. These two ratings for each reported collision were summed producing a risk index for each experienced fire fighter that combined frequency of collision, severity and blame. The mean number of reported collisions were 0.56 and 2.85 for low and high-risk groups, with mean summed severity/blame scores of 1.7 and 10.7, respectively. A split of participants based on their risk indices resulted in 23 participants classified as low-risk (on or below the median) and 20 participants considered high-risk (all above the median).

The final group was made up of 20 control drivers (19 male, 1 female). Their mean age was 43.9 years, with 9252 personal miles per annum, and they had a mean personal driving experience of 22 years since passing their driving test. A comparison of age and personal driving experience between the control group and the fire fighter cohort as a whole did not reveal any significant differences ($p > 0.1$).

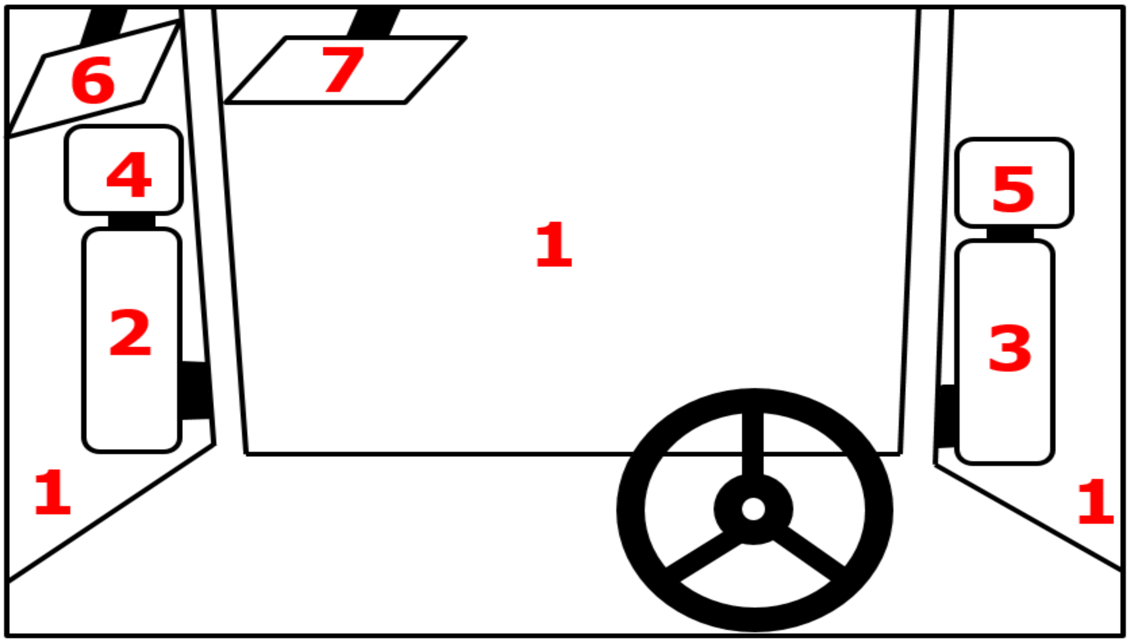
Materials and apparatus

Filming

The fire-appliance hazard perception test was developed from footage that was captured from multiple fire appliances on blue-light training runs. All clips were filmed around Nottinghamshire over a four-week period in April – May 2015. The filming took place during a number of Emergency Fire-Appliance Driver (EFAD) courses to avoid the necessity of undertaking additional non-emergency blue-light runs beyond those required for training purposes. In total approximately 12 hours of footage was obtained from the fire appliances.

Filming from the fire appliances required a 7 camera system in order to capture the forward view from the cabin and the 6 views that are available to the driver through the mirrors (See Figure 1a to see a schematic representation of the separate video feeds). The mirror information was subsequently combined with the forward view, and with a graphic overlay of the cabin interior to create an immersive experience (see Figure 1b for a screen shot from a finished clip).

(A)



(B)

Figure 1. Panel A: A schematic depiction of the envisioned view of the final edited clips, with numbers relating to the different video feeds; Panel B: a screen shot from one of the final fire appliance clips that combines all seven video feeds with the graphic overlay of the cabin interior.

A GoPro HERO4 Silver Edition camcorder recording in Full High Definition format (1080p, 16:9 ratio, wide-angle setting) was positioned on the dashboard of the fire appliance to capture the forward view. For mirror views, six JVC Action Cameras (Model Number: GC-XA1BU; 1080p, 16:9 ratio) were mounted externally using suction mounts aligned with the mirrors, but positioned to avoid obstruction for the driver. Four of these cameras were mounted on the doors to capture wing mirror and blind spot mirror views (feeds 2, 3, 4, & 5 in Figure 1a). One further camera was positioned on the left of the vehicle pointing downwards to provide kerb distance information (feed 6), with a final camera placed on the external windscreen pointing downwards to capture the blind spot in front of the cab (feed 7). All external cameras were tethered to the vehicle for safety.

Creating the tests

Prior to video editing, a graphic overlay was designed to represent the interior of a fire appliance (see Figures 1b). A-pillars and the internal roof of the fire-appliance cabin was designed to be partially transparent to prevent these parts of the graphic overlay from obscuring aspects of the forward view. This was done to mimic the effects of stereopsis and head movements, which naturally minimise A-pillar obscuration in real driving.

Footage from the multiple cameras was synchronised in Adobe Premiere CC, and then reviewed by a team of transport psychologists and fire service personnel in order to select the most promising stimuli. A total of 30 clips were chosen on the basis that they provided at least one hazard of sufficient concern to warrant a response. These hazards also had to have precursors (i.e. a non-hazardous element of the scene that foreshadows a potential hazard. Such precursors are essential for a hazard perception test as they provide subtle cues that prime the impending hazard, which safer drivers are more likely to spot and comprehend than less-safe drivers. Hazard onset times (i.e. the earliest point at which participants could make a correct response to the hazard) were based on the point at which an obstacle begins to move into the path of the approaching fire appliance. Hazard offsets (i.e. the latest point at which a participant could make a correct response to the hazard) reflected the point at which a response would no longer beneficial to helping avoid the hazard. A description of the selected hazards is given in Table 1. The clips did not contain an audio track.

The thirty clips were divided into two tests each containing 15 clips. Half of the drivers saw clips 1-15 as a hazard perception test (while clips 16-30 were presented as a hazard prediction test: see experiment 2), and the other half of the participants viewed clips 16-30 as a hazard perception test (and clips 1-15 as a hazard prediction test).

Data collection apparatus

The hazard perception test presented on a computer monitor, measuring 48.3 cm x 30.5 cm. The monitor was connected to a SensoMotoric Instruments' Remote Eye-tracking Device, sampling at 500Hz (SMI RED 500) with a 50 ms threshold for fixations. Participants were provided with a keyboard to make speeded hazard responses.

Design

A 1x4 between-groups design was employed, with four driver groups: control drivers, novice fire appliance drivers, high-risk, experienced fire-appliance drivers, and low-risk, experienced fire-appliance drivers. All participants watched 15 hazard perception clips, presented in a random order, and were required to press a button on a keyboard to indicate that they had detected a hazard. Each hazard contained one *a priori* hazard that was chosen in consultation with Fire Service Driving Instructors. Response times to these hazards were the primary dependent variable.

Responses were considered valid if they fell within a specific temporal hazard window, defined by the hazard onset and offset points for each clip. Hazard response times were calculated from the hazard onset.

329 *Table 1.*

330 A description of the hazards in the final 30 clips selected for the Fire Appliance Hazard
331 Perception test (onsets and offsets refer to the HPT).

Clip no.	Hazard	Description	Hazard onset (ms)	Hazard offset (ms)
1	Car remains stationary in the road ahead.	The fire appliance is travelling on a 30mph urban road. Ahead, a lollipop lady is in the road allowing children and pedestrians to cross. A car is waiting at the lady preventing the appliance from making progress.	23134	30634
2	Pedestrian in the road.	The fire appliance is travelling on the tram tracks. A pedestrian, hidden from view by other pedestrians on the pavement, enters the road in front of the appliance.	39967	42900
3	Workman in the road.	The fire appliance is travelling on 30mph suburban road. A workman, partially obscured by a work lorry, is working in the road and does not notice the appliance.	25067	27034
4	Large lorry ahead.	The appliance is travelling on a 30mph inner city road. The appliance approaches a set of traffic lights of which the left-side turn and view is blocked by a large building. As the appliance approaches, a large lorry from the left pulls out in front of the appliance.	21000	26534
5	Van with trailer pulls out.	The appliance is travelling down a 30mph road. A van towing a trailer does not notice the appliance and pulls out in front of it to overtake a car that has pulled over on the left.	25200	29900
6	Pedestrians walk in the road.	The appliance is travelling down a 30mph road. A mother with children turns on the right-hand pavement begins to cross. She notices the appliance and stops in the road.	32100	34300
7	Van pulls out.	The appliance is travelling down a 30mph road. As the appliance approaches a small road island, a large bus blocks the right-hand view and a van from the right pulls out in front of the appliance.	22900	25800
8	Car pulls out.	The appliance is travelling down a 30mph inner city road. As the appliance approaches a set of traffic lights, the traffic coming from the right has their view	34367	38100

		restricted by a building, as such, a car does not see the appliance and pulls out in front of it.		
9	Car pulls out .	The appliance is travelling on a 30mph road. As the appliance approaches a set of traffic lights, the traffic coming from the left have their view restricted by housing. A car does not see the appliance and pulls out in front of it.	33034	35634
10	Car pulls out.	The appliance is travelling around a roundabout with traffic lights. A car from a left-hand side road does not notice the appliance and pulls out directly in front of it.	43434	48767
11	Large lorry ahead.	The appliance is travelling down a narrow urban road. Ahead is a set of traffic lights with both the left and right-side views blocked by buildings. As the appliance approaches, a large lorry from the right turns, partially blocking the road.	20100	35767
12	Pedestrians walk in the road.	The appliance is travelling down a 30mph road. Pedestrians from the right-hand pavement begin to walk into the road.	21300	23234
13	Van pulls out.	The appliance is travelling down a 30mph road. Ahead there is a bend in the road to the left. As the appliance approaches the bend, a van on the opposite side of the road (hidden by the bend) turns directly in front of the path of the appliance.	27234	31900
14	Car almost pulls out.	The appliance is travelling down a 30mph road. Ahead, a large car from a right-hand side street almost pulls out in front of the appliance.	15167	18334
15	Mobility scooter pulls out.	The appliance is travelling down a 30mph road. The road begins to incline, just past the brow of the hill, a mobility scooter enters the road from the right, directly in front of the appliance.	27167	31967
16	Pedestrian in the road.	The appliance is travelling down a 40mph road. As the appliance approaches a set of traffic lights, a pedestrian is walking in the middle of the road.	5967	14634
17	Pedestrian in the road.	The appliance is travelling down a 30mph road. A pedestrian hidden from view by a lorry parked on the left-hand side of the road enters the road and crosses in front of the appliance.	43767	46400
18	Car almost pulls out.	The appliance is travelling down a busy 30mph road. A car, hidden from view by the stream of traffic on the	32634	37234

		opposite side of the road, almost pulls out of a right-hand side road.		
19	Pedestrians walk in the road.	The appliance is travelling down a 30mph road. Pedestrians hidden from view by queuing traffic on the right-hand side of the road, enters the road and crosses in front of the appliance.	37300	42200
20	Stabilising leg of work lorry blocks road.	The appliance is travelling down a 30mph urban road. Ahead, a large work lorry with a stabilising leg restricts the road, turning it into a single carriage.	31234	39867
21	Car reverses towards appliance.	The appliance is travelling down a 30mph road. Ahead, a car waiting at the traffic lights begins to reverse towards the appliance.	25367	31767
22	Ambulance on blue lights invades lane.	The appliance is travelling down a 30mph road approaching a pedestrian crossing. Ahead, an ambulance car on blue-lights overtakes the traffic waiting at the pedestrian crossing and invades the lane the appliance is in.	23234	31434
23	Car pulls out.	The appliance is travelling down a 30mph road with two lanes. The lane on the right has heavy queuing traffic. A car in this lane does not see the appliance and suddenly pulls out of the busy lane directly in front of the appliance.	40800	47534
24	Car pulls out.	The appliance is travelling down a 30mph road. As the appliance approaches a traffic-light controlled cross roads, the right-hand view is blocked by a large building. A van coming from the right, turning left, stops in the road but unintentionally blocks the view of the appliance from other road users. A car from the right, going straight ahead, pulls out from behind the van, directly in front of the appliance.	40567	42300
25	Pedestrian in the road.	The appliance is travelling down a 30mph road. A pedestrian on their mobile phone steps into the road from the left-hand side pavement in front of the appliance.	27034	30700
26	Cyclist veers towards appliance	The appliance is travelling down a 30mph road. A cyclist on the right-hand side of the road veers towards the appliance.	29000	30967

27	Ambulance encroaches on the lane	The appliance is travelling down a 30mph road. The appliance approaches a set of traffic lights and turns right, as the appliance turns, an ambulance on blue-lights on the opposite side of the road approaches, invading on the appliance’s lane.	20000	24000
28	Car pulls out.	The appliance is travelling down a 30mph road. Ahead, a car parked on the left-hand pavements pulls out in front of the appliance.	27234	31700
29	Car pulls out.	The appliance is travelling down a 30mph road. Ahead, a car from the left-hand side road pulls out in front of the appliance.	15034	22700
30	Pedestrian almost walks out.	The appliance is travelling down a 30mph road. As the appliance approaches a pedestrian crossing a pedestrian almost walks out in front of the appliance.	8034	10234

Additional measures included the percentage of *a priori* hazards responded to, and a selection of eye movement measures (time to first fixate the hazard, first fixation duration on the hazard, mean fixation duration on the hazard, number of fixations on the hazard, and total dwell time on the hazard). All response and eye movement data were only considered to relate to the hazard if they occurred during the hazard window (i.e. the period of time between hazard onset and hazard offset). Additionally, eye movements during the hazard window had to fall directly upon the hazard (+ approximately 1 degree of visual angle) to be considered as relevant fixations. These measures were analysed primarily via a series of 1x4 Analyses of Variance (ANOVAs) comparing across the four participant groups.

Procedure

Fire Service personnel were tested in a quiet office in their respective Nottinghamshire fire stations while on shift. Control participants were tested within an eye tracking laboratory at Nottingham Trent University. Each participant was first asked to complete a battery of questionnaires: demographics, driving history, the Driver Behaviour Questionnaire (DBQ; Reason et al., 1990; Parker et al., 1995), the Traffic Locus of Control (T-Loc; Özkan and Lajunen, 2005), and the Sensation Seeking Scale (SSS; Zuckerman, 1976).

Participants undertook 3 tests in total: the hazard perception test (experiment 1), the hazard prediction test (see experiment 2), and a third test based on gap judgements (this latter test is not discussed in the current paper). The order of the perception and prediction tests was counterbalanced, and they were presented either before or after the gap judgement task.

Participants were seated approximately 60cm from the screen and told that they would see video clips taken from the perspective of a fire-appliance driver, driving in an emergency response situation (i.e., a blue-light run). They were instructed to press a button as quickly as possible to indicate the presence of a hazard that would require them to suddenly stop, slow down or change position in some way to avoid a potential collision. All participants saw a practice clip before beginning the experiment.

Results

One-way Analyses of Variance (ANOVA) compared the four groups across a range of measures for the hazard perception test. Following the omnibus analyses a series of planned Helmert contrasts were conducted. These sub-analyses compared (a) the scores of control participants to the mean scores of all fire-appliance drivers, (b) the scores of novice, FA drivers to the mean scores of all experienced, FA drivers, and (c) the scores of high-risk, experienced, FA drivers to those of low-risk, experienced, FA drivers. These contrasts reflect the sub-hypotheses for the study: all FA drivers should out-perform control drivers; all experienced, FA drivers should out-perform novice FA drivers; and low-risk, experienced FA drivers should out-performance the high-risk FA drivers. Any significant contrast effects were adjusted for potential familywise error via Hochberg corrections, with differences accepted at the 0.05 level for 1-tailed tests (reflecting the directional nature of the *a priori* predictions).

Response times

One low-risk, experienced, FA driver was removed from the analysis as the number of hazards he detected fell more than 3 standard deviations below the mean detection rate for the whole sample. Response times (RTs) were calculated from the *a priori* hazard onset times. Failures to respond to a hazard were assigned a maximum response time, equivalent to the hazard offset (following McKenna et al., 2006). To minimize skew in the data a square root transform was used. The transformed RTs were then standardised into Z-scores using the overall sample mean and standard deviation (SD) for each hazard. This process was necessary because the hazard windows varied in duration, and without standardisation, some hazards might exert a greater influence on the final mean score than others (following Wetton et al., 2010). While all analyses were conducted on these z-scored, square-root transformed RTs, for clarity of presentation in graphs these figures were converted back into millisecond response times using the mean and standard deviation across all hazards and participants. The converted response times for the four participant groups appear in Figure 2.

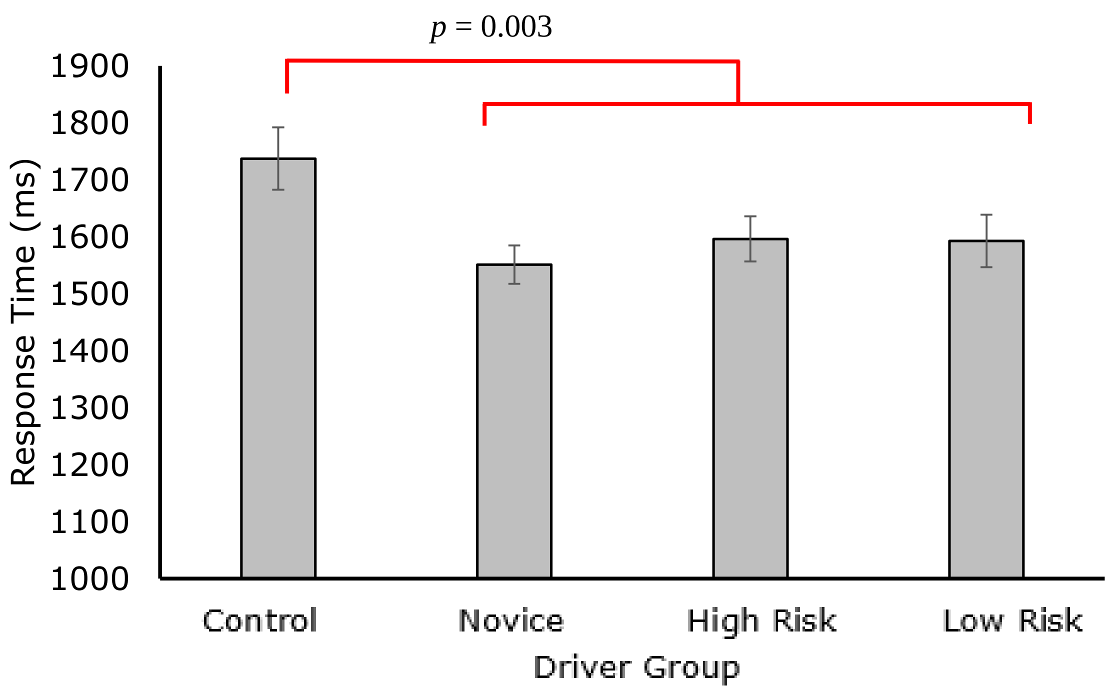


Figure 2. Response time to hazards (ms) across the four participant groups (with standard error bars added). Note: these scores have been converted back from Z-scores.

A 1 x 4 between-subjects ANOVA on the response time data revealed a main effect of driving experience, $F(3, 79) = 3.35$, $MSe = 0.48$, $p = 0.02$. Planned Helmert contrasts confirmed that control drivers were slower to detect the hazard than all other fire-appliance driver groups (1737 ms vs. 1580 ms; $p = 0.003$). There were no differences between the three groups of fire-appliance drivers (all *p*s > 0.05).

A similar 1 x 4 ANOVA was conducted on the percentage of *a priori* hazards that participants responded to (control = 77%, novice = 85%, high-risk = 79%, low-risk = 83%). The omnibus effect was not significant, and none of the planned contrasts reached significance.

Eye movement measures

The first analysis compared the percentage of *a priori* hazards that participants fixated (at least one fixation on the hazard, between onset and offset). Though the omnibus calculation was not significant ($F(3, 79) = 1.89$, $MSe = 166.61$, $p = 0.40$), the planned Helmert contrasts revealed a significant difference between novice fire-appliance drivers and experienced fire-appliance drivers suggesting that the experienced drivers looked at more hazards than the novices (90.7% vs 85.0%, respectively; $p = 0.04$; see Figure 3). Following correction for familywise error, this comparison was marginal at best ($p = 0.057$).

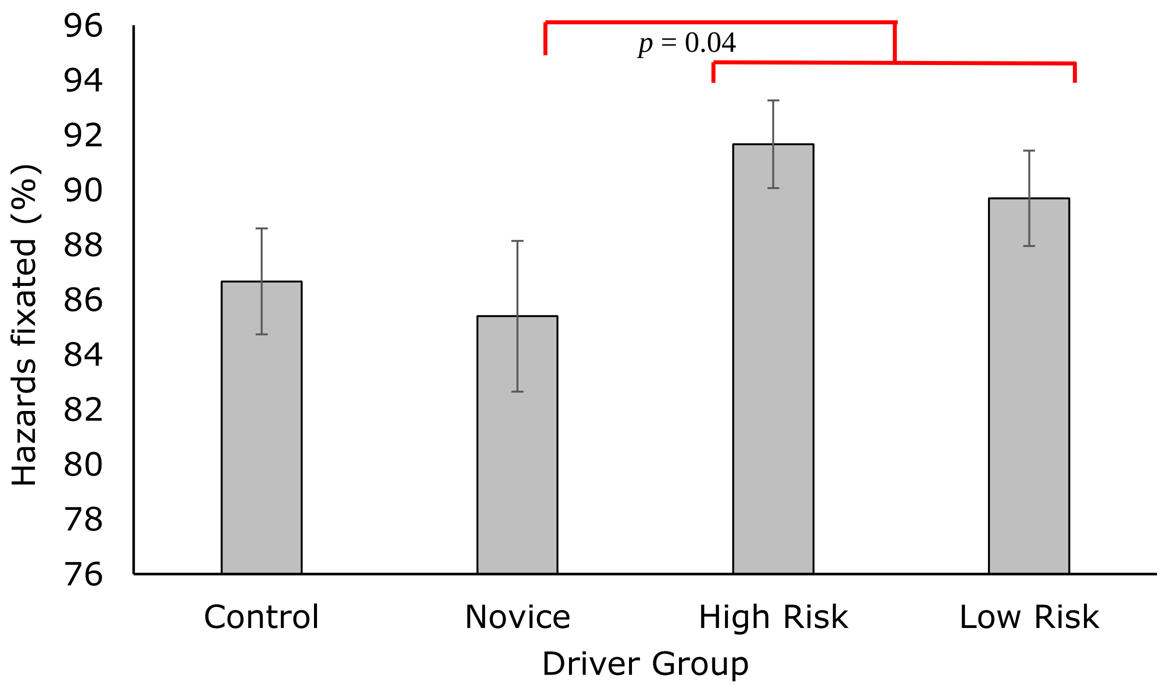


Figure 3. The percentage of hazards that participants fixated at least once, across the four driver groups (with standard error bars added).

The number of hazards that were fixated was high, reflecting the fact that as the hazard window progresses, the hazards become more obvious and more likely to attract attention. Thus, a more sensitive measure might be the time taken to first fixate the hazard following onset. For this analysis, if a participant was looking at the appropriate location on the screen at the point of hazard onset, as if they had successfully predicted that a hazardous precursor would develop into a full hazard, they were considered to have a time-to-fixate latency of 0 ms. If, however, drivers failed to look at the hazard during the hazard window, they were given the maximum time possible, equivalent to the hazard offset (following McKenna et al., 2006). These measures were square-root and z-score transformed in order to reduce skew and ensure comparability across clips (as with the response times).

A 1 x 4 between-subjects ANOVA revealed a significant main effect of driver experience, $F(3, 79) = 4.95$, $MSe = 0.55$, $p = 0.03$. Planned Helmert contrasts identified control drivers as slower to fixate the hazards than all fire-appliance driver groups ($p = 0.03$), though this appears to be driven by the short fixation latencies of the two experienced fire-appliance groups, who were also faster to fixate than the novice fire-appliance drivers (831 ms vs. 960 ms, respectively; $p = 0.003$; see Figure 4). There was no difference between high and low experienced fire-appliance drivers in terms of how quickly they fixated the hazards.

Several measures were recorded to reflect the amount of attention that participants gave to the hazards. These included first fixation duration (the length of the first fixation given to a hazard by a participant), mean fixation duration (the average duration of all fixations given to each hazard), the number of fixations on each hazard, and the dwell time on hazards (the number of eye tracking samples that fell on the hazard during the hazard window, z-scored for comparability across clips). All of these measures were compared across the four driver groups, but no significant differences were found.

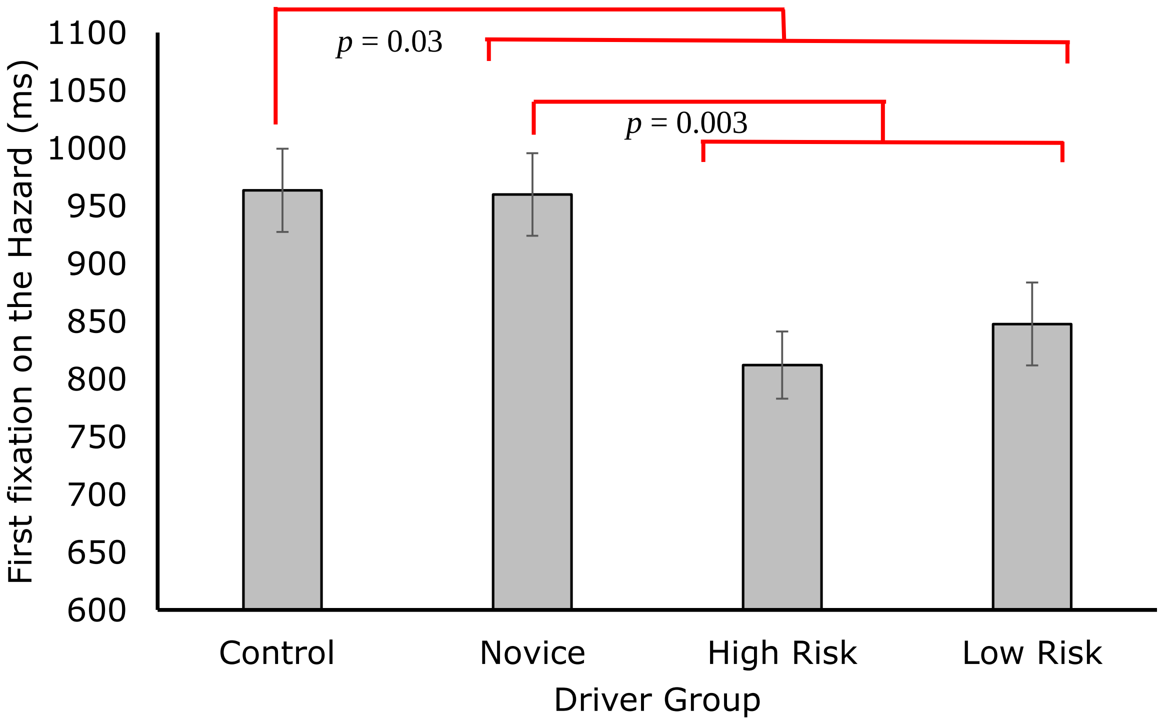


Figure 4. The average time taken to fixate the hazard for each driver group (with standard error bars added). Note: these scores have been converted back from Z-scores.

In addition to measures of attention devoted to the hazard, we also calculated the amount of time devoted to the hazard precursor. A precursor typically precedes a hazard and acts as a clue to the upcoming hazard. For instance a pedestrian on the pavement walking towards the road, may lead to the prediction that the same person may step out into the road and become a hazard. Measures of attention to these precursors reflect the preparatory work that drivers undertake in actively predicting imminent hazards.

For the current analyses, the measure of dwell time was chosen to reflect attention given to the hazard precursors. The precursor was defined as the most appropriate clue to the hazard, and was typically located in the same physical space as the actual hazard, but preceded it in time (on many occasions the precursor was the hazardous object, but before it became hazardous). The dwell-time measure was calculated as the sum of all eye-tracking samples that fell on these precursors in a 1000 ms time window immediately preceding the hazard onset. By using a set temporal window, we did not need to convert dwell times to z-scores.

A 1 x 4 between-subjects ANOVA was conducted on the precursor dwell times. This revealed a marginally significant effect of driving experience ($F(1,79) = 2.7$, $MSe = 5158$, $p = 0.05$). Helmert contrasts demonstrated that novice fire-appliance drivers were likely to have significantly less dwell on the hazard precursors than experienced fire-appliance drivers (149 ms vs. 195 ms, $p = 0.02$; see Figure 5).

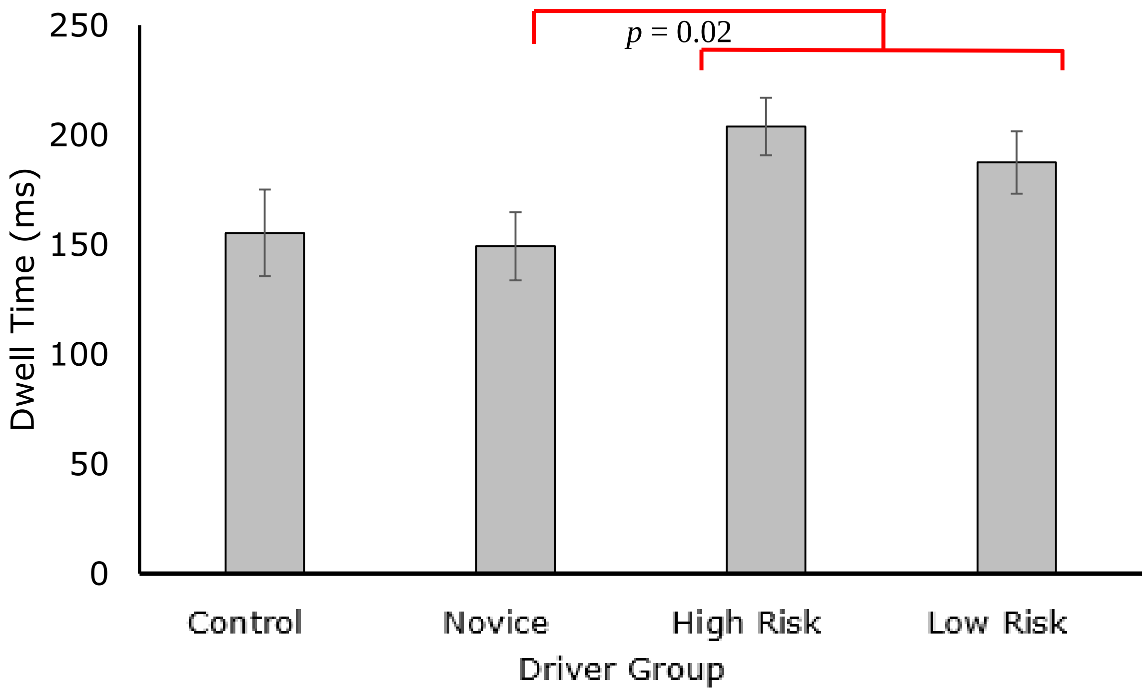


Figure 5. The average dwell time (ms) on the precursor across the different participant groups (with standard error bars added).

Questionnaire measures

Of all the questionnaire measures taken, only the Driving Behaviour Questionnaire (Reason et al., 1990; Parker et al., 1995) proved interesting. Twenty-four items were given, split into 3 factors: violations, errors, and slips/lapses. Cronbach's alpha for all three was acceptable (0.83, 0.73, 0.66, respectively).

The resultant participant means for the three factors were entered into a series of 1 x 4 ANOVAs. In the analysis of errors, the omnibus test was not significant, $F(3, 79) = 2.14$, $MSe = 0.50$, $p = 0.10$, however planned Helmert contrasts revealed that low-risk, experienced fire-appliance drivers scored significantly lower on the error factor of the DBQ (i.e. reported fewer errors) than high-risk, experienced fire-appliance drivers (1.47 vs. 1.82; $p = 0.02$). No other contrasts reached statistical significance (all ps > 0.05).

The omnibus test on scores for the violation factor was also non-significant ($F(1, 79) = 2.23$, $MSe = 0.92$, $p = 0.09$), but the planned contrasts revealed that low-risk experienced fire-appliance drivers reported significantly fewer violations than the high-risk drivers (1.60 vs. 2.08; $p = 0.02$). No other contrasts reached statistical significance (all *p*s > 0.05).

Finally, the omnibus test for slips and lapses also struggled to reach significance ($F(1, 79) = 2.34$, $MSe = 0.59$, $p = 0.08$), but the contrasts once again revealed low-risk experienced fire-appliance drivers to report fewer lapses than the high-risk drivers (1.89 vs. 2.22; $p = 0.04$). Following correction for familywise error however, this comparison was marginal at best ($p = 0.057$). No other contrasts reached statistical significance (all *p*s > 0.05).

Discussion

To summarise the results, all fire-appliance drivers responded faster to hazards than the control group, though there were no differences between the groups of fire-appliance drivers. The two experienced, fire appliance groups were, however, more likely to look at the *a priori* hazards. Novice fire-appliance drivers looked on average at 85% of the hazards, and responded to 85%, whereas the experienced fire-appliance drivers looked at 91% of hazards on average, yet only responded to 80% (which does not differ significantly from the mean novice response rate). We therefore suggest that both of the experienced groups were potentially aware of more potential hazards, yet decided to only respond to a proportion of those that they looked at (albeit a high proportion).

The experienced FA drivers were also noted to fixate the hazards sooner than the novice drivers (see Crundall et al., 2012 for similar results with driving instructors in a simulator; cf. Huestegge et al., 2010, who failed to find such an effect when using static images). Our experienced drivers were also found to spend more time looking at the precursors to the hazard. Together these results provide a clear story: the experienced FA drivers are better able to anticipate hazards. They spend more time looking at the precursors (or clues) to imminent hazards, suggesting that they can effectively prioritise those areas and objects within the scene that may give rise to a hazard. Through their prioritisation of these precursors, the experienced drivers are more likely to spot when a precursor turns into an actual hazard. This is reflected in their speed to fixate hazards and their higher proportion of hazards fixated overall. There was no difference between the high-risk and low-risk groups on any measure however, suggesting that either hazard perception skill is not relevant to their risk level, or that the test was not sensitive enough to evoke and record risk-related differences in behaviour in response to the hazards.

The homogeneity of response times across the three fire appliance groups can be explained in two ways. First the experienced FA drivers may be applying a higher threshold for what they consider to be a hazard. This has been found previously with police drivers (Crundall et al., 2003) and may reflect their self-perception of driving skill (i.e. experienced drivers are more likely to look at the hazard and think ‘It may be a hazard, but I could handle it’ and therefore be less likely to press the button to acknowledge it. This is supported by the disparity between the number of hazards fixated and the number responded to by experienced drivers).

Secondly, it may be the case that novice FA drivers have been sufficiently trained to be able to respond to on-road hazards with very quick responses. Even though they are slower to look at these hazards, when they finally do look at them, their training may allow rapid processing leading to a quick response. While this explanation might reflect the success of the training undertaken by the novice drivers, it still suggests that novice drivers have not yet developed the anticipatory skills that the more experienced drivers demonstrate.

Previous studies have also found eye movement differences between groups that have not translated into response differences (Chapman and Underwood, 1998; Crundall et al., 1999). This suggests that the stimuli are sufficient to provoke experiential differences in behaviour, but that the simple response-time measure of the traditional hazard perception test maybe too insensitive to detect them. Unfortunately, a test of hazard perception skill must ultimately rely on simple behavioural measures (rather than eye movements or physiological responses) in order to achieve wide-spread take-up by the fire service.

There are, however, a number of ways to iterate the test in order to obtain a simple response time measure that better reflects the underlying eye movement differences between novice and experienced fire-appliance drivers. First, more detailed instructions could be provided to participants regarding the decision to make a response to the hazard. By providing more concrete examples of desired hazard responses, we would hope to convert some of the hazards that experienced drivers spotted but decided not to report, into positively identified targets . At the same time, it could be useful to clearly define hazards not as things that ‘you would have to brake suddenly for’, but as things that ‘an average driver would have to brake suddenly for…’. This approach may also encourage experienced drivers to respond to hazards that they feel eminently capable of handling themselves, but which they acknowledge might be difficult for less-experienced drivers.

Secondly, a traditional method of titrating clips is to analyse them individually to identify whether there are any clips that are extremely poor indicators of group differences. By removing specific clips we can then pare the test down to only include those clips that most clearly discriminate between experienced and novice drivers. Ideally, this would involve undertaking the initial study with a much wider range of clips, though the practicalities of collecting more footage and conducting longer studies with on-duty fire fighters prevented this.

Finally, we may try a different approach all together. An alternative variant on the traditional hazard perception test was proposed by Jackson et al (2009). Initially termed the 'What Happens Next?' test, this targets the sub-component of hazard prediction skill, arguably the most important of the hazard perception sub-skills. When measured in isolation it can provide an ostensibly more robust discrimination between safe and less-safe driver groups, unconfounded by the multiple underlying sub-processes that afflict the traditional hazard perception measure. It is for this reason that we designed a hazard prediction test which was run concurrently with the hazard perception test. The results of the hazard prediction test are presented in the following sections.

Experiment 2

The second experiment is based on the occlusion technique first used by Jackson et al. (2009), and expanded upon by several subsequent studies (e.g. Castro et al., 2014; Crundall, 2016; Lim et al., 2014; Ventsislavova et al., submitted). Each video ends abruptly as the hazard begins to develop and the scene is occluded.

Jackson et al. (2009) demonstrated that occlusion is necessary to discriminate between experienced and novice drivers, as the alternative of leaving a frozen image of the final frame allowed novices additional time to seek out the answer. Thus the successful driver presumably needs to be looking at the right place at the right time (and probably be expecting the right thing to happen) in order to see the hazard. Drivers who successfully predict the upcoming hazard will have an advantage in this regard.

The choice of occlusion point is ostensibly of vital importance. If one cuts the clip too late, everyone sees the hazard: no prediction is needed, and no discrimination will be found between safe and less-safe drivers due to a ceiling effect. Equally however, if one cuts the clip too early, without any possible clue to the upcoming hazard, then a floor effect will remove group differences. In-between these two extremes however, minor variations in the occlusion point appear to have little effect on the discriminability of the test (Crundall, 2016). While earlier occlusions reduce the overall number of drivers who correctly predict the hazard, discrimination between novice and experienced drivers is maintained providing that some clue to the impending hazard remains.

In the current study we opted to occlude mere hundreds of milliseconds after hazard onset. The rationale for ending the clip just after hazard onset is that the handful of video frames containing the initial development of the hazard gives the participant confirmation that their prediction is correct. The briefness of this post-onset event is so slight however, that it is unlikely to be registered by anyone who is not already looking at the appropriate location.

The current experiment also follows the innovation of two studies (Castro et al., 2014; Lim et al 2014) in providing multiple-choice answers. Other studies (Jackson et al., 2009; Crundall 2016, Ventsislavova et al., submitted) have required verbal or written predictions from participants. While these provide rich data, this method is reliant on subjective coding and cannot be automatically marked to provide an immediate score. For this study we have followed the more pragmatic testing approach of providing 4 options, with one correct answer embedded in 3 distracter answers.

The hypotheses for this experiment remained the same as that for experiment 1: all fire service personnel will out-perform controls, experienced FA drivers will out-perform novices, and low-risk, experienced drivers will out-perform high-risk, experienced drivers.

Method

The same participants from experiment 1 undertook the current study, split into control drivers, novice FA drivers, low-risk, experienced FA drivers and high-risk, experienced FA drivers. Experiment 1 and 2 were counterbalanced across participants within the testing session.

The methodology of experiment 2 is identical to that of experiment 1, except for the following modifications. The clips from experiment 1 (see Table 1) were edited to finish just as the hazard began to develop or become visible. A precursor to the hazard was always available, though the duration of precursors varied across the clips. At the point of occlusion, a screen was immediately presented displaying the question 'What happens next?'. Four options were also provided, and participants were required to choose the most likely answer. Both the correct answer, and suitable distracter options, were decided in discussions between a group of transport psychologists and fire service personnel. Distracters were chosen that were as feasible as possible given the available precursors in each given scene, and were chosen on the basis of consensus. The order of the correct answer and the three distracter options on the screen was randomly determined for each clip. Participants were required to select the most appropriate answer using a computer mouse. They were aware that selection of the answer was not timed.

The main dependent variable for this test was participant percentage accuracy in choosing the correct option across 15 clips. Other measures included the time to first fixate the hazard precursor, first fixation duration on the precursor, mean fixation duration on the precursor, number of fixations on the precursor and total dwell time on the precursor. Fixations were considered to have landed on the precursor if they occurred during the prediction window leading up to occlusion, and were spatially located on the actual element of the scene that acted as the precursor to the hazard (i.e. the clue to the imminent danger + approximately 1 degree of visual angle). As the precursor was the only relevant stimulus that could be fixated, these windows were tailored to the natural duration of the precursor, rather than using a shorthand 1 second window as in Experiment 1. Prediction windows began when the clue to the hazard was first visible (e.g. a pedestrian becomes visible on the pavement) and ended when the hazard has just started to develop (typically 150 to 250 milliseconds after hazard onset, as defined in table 1).

It was predicted that all driver groups would differ, with FA experience and low-risk leading to better prediction accuracy, underpinned by group differences in participants' eye movements. Given recent evidence (Ventsislavova et al., submitted), we expected the prediction test to provide stronger discrimination between the groups than the perception test used in Experiment 1.

Results

One-way Analyses of Variance (ANOVA) compared the four groups on their percentage accuracy in the prediction task, and on a range of eye movement measures. Planned Helmert contrasts were again conducted to assess differences between controls and all FA drivers, between inexperienced and all experienced FA drivers, and between two groups of experienced FA drivers split according to risk. The poorly performing outlier identified in Experiment 1 (a low-risk, experienced fire-appliance driver) was also removed from the current analysis for the sake of parity across studies. This was a conservative decision, as his performance on the prediction study was much better than on the initial study.

Prediction accuracy

When the percentage accuracies for all participants were compared in a 1 x 4 ANOVA a main effect of driving experience was revealed, $F(3, 79) = 2.93$, $MSe = 382.48$, $p = 0.04$. Planned

Helmert contrasts revealed that all fire-appliance drivers were significantly more accurate at predicting upcoming hazards than matched controls (69.2% vs. 63.3%, respectively; p = 0.05). It was also noted that high-risk, experienced fire-appliance drivers scored similarly to the novice drivers, and were therefore significantly worse at the prediction test compared to the low-risk driver group (65.3% vs. 73.0%, respectively; p = 0.03; see Figure 6).

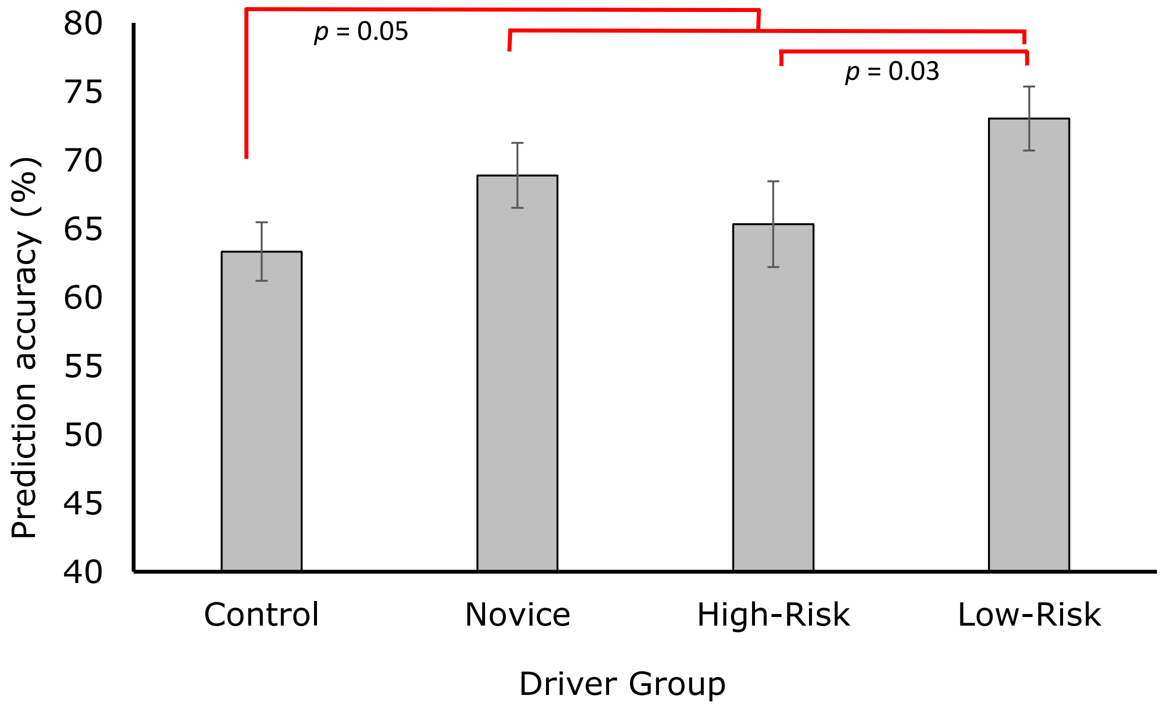


Figure 6. The mean prediction accuracy (%) across the four driving groups for the 'What Happens Next' test (with standard error bars).

Eye movement results

The eye movement data of four further participants were removed due to loss of calibration during the test (one novice FA driver, one low-risk, experienced driver and two control drivers).

Participants did not have much opportunity to look at the actual hazards in the prediction test, as the screen would occlude just as the hazard would begin to unfold (mere hundreds of milliseconds following hazard onset, as defined in Table 1). However any fixations that fell within the temporal prediction window upon the hazard precursor (+ 1 degree of visual angle approximately), were considered to reflect how safer drivers can predict and seek out hazards before they occur.

The first analysis of eye tracking data on the prediction test merely compared the percentage number of clips during which the drivers fixated the precursor within the prediction window. When subjected to a 1 x 4 between-groups ANOVA, this revealed a main effect of driving experience, $F(3, 75) = 4.06$, $MSe = 880.51$, $p = 0.01$. Planned Helmert contrasts showed that control drivers fixated significantly fewer precursors than all fire-appliance drivers (48.89% vs. 60.33%, respectively; $p = 0.005$). There was a suggestion in the means that low-risk fire-appliance drivers might fixate more precursors than high-risk fire-appliance drivers, but this difference did not reach conventional levels of statistical acceptability (65.4% vs. 57.7%, respectively; $p = 0.09$; see Figure 7).

The *time to first fixate* hazard precursors was calculated as the start of the first fixation within the prediction window that landed on the hazard precursor, minus the time at which the prediction window opened for each clip. If participants did not look within the prediction window prior to occlusion they were assigned the maximum possible time to fixate (i.e. the full length of the prediction window; following McKenna et al.'s treatment of missing RT values, 2006). If participants were already looking at the appropriate location when the prediction window opened, they were given a *time to first fixate* of zero milliseconds. These measures were square-root and z-score transformed in order to reduce skew and ensure comparability across clips. Although the pattern of results followed that found in Figures 6 and 7, with low-risk experience drivers having the shortest time-to-fixate, and control drivers taking the longest to fixate the precursor, the main effect did not reach significance ($F(3, 75) = 2.14$, $MSe = 0.10$, $p = 0.10$).

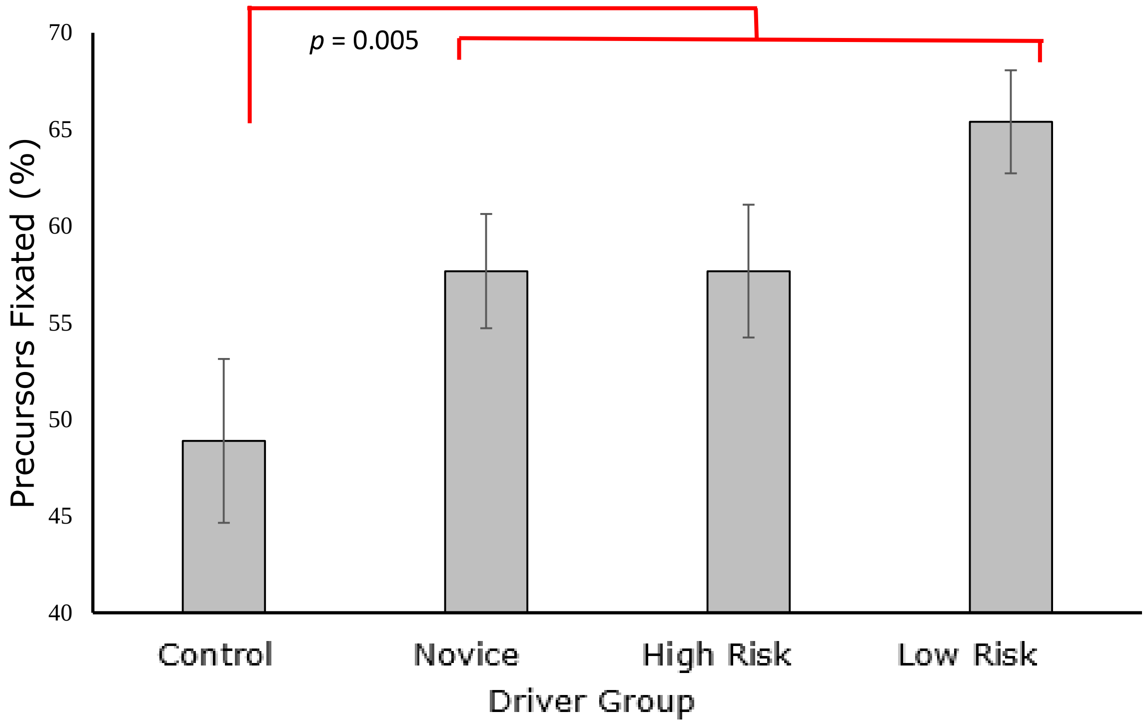


Figure 7. The average percentage of hazard precursors that were fixated for each driving group (with standard error bars).

While the time to first fixate the hazards in the hazard perception test (Experiment 1) is an informative measure that tells us which group of participants spot the hazard soonest, it is arguable how useful this measure is in the case of precursors in the current prediction test. When the precursor first becomes visible it contains very little information, and fixations upon precursors at this point may not reflect the meaningful extraction of hazard evidence (Crundall et al., 2012; Pradhan and Crundall, 2017). As the clip progresses, the precursor becomes more informative, with the most informative point being just before hazard onset. Therefore in order to predict what happens next, we might expect that the most accurate responders will be those who are looking at the precursor at the very moment that it changes into a hazard, just as the screen occludes (i.e. the safest drivers should have the smallest temporal gaps between last fixating the precursor and the onset of the hazard). On this basis we suggest that the temporal proximity of the last fixation on the precursor to the occlusion point is more important than the first fixation on the precursor.

To assess this hypothesis, the occlusion point for each hazard was subtracted from the end point of each participants' final fixation within the prediction window, providing a measure of *last-precursor-fixation-to-hazard lag*. If participants did not look within the prediction window prior to occlusion they were assigned the maximum possible lag (i.e. the full length of the prediction window; following McKenna et al.'s treatment of missing RT values, 2006). If participants were however looking at the appropriate location at the point of occlusion, they were given a lag of zero milliseconds.

A 1 x 4 between-groups ANOVA on these data revealed a main effect of *last-precursor-fixation-to-hazard lag* ($F(3,75) = 5.70$, $MSe = 0.01$, $p = 0.001$). Planned Helmert contrasts revealed that control drivers had a greater lag than all fire-appliance drivers (i.e. they were less likely to be looking at the precursor at the time of occlusion; 719ms vs. 635ms, p = 0.001), and that high-risk, experienced drivers had a greater lag than low-risk, experienced drivers (667ms Vs. 600ms, p = 0.02). As can be seen from Figure 8, the low-risk, experienced fire-appliance drivers were fixating the precursor at the closest point to the occlusion on average, suggesting they were the group most likely to be expecting the appearance of the hazard.

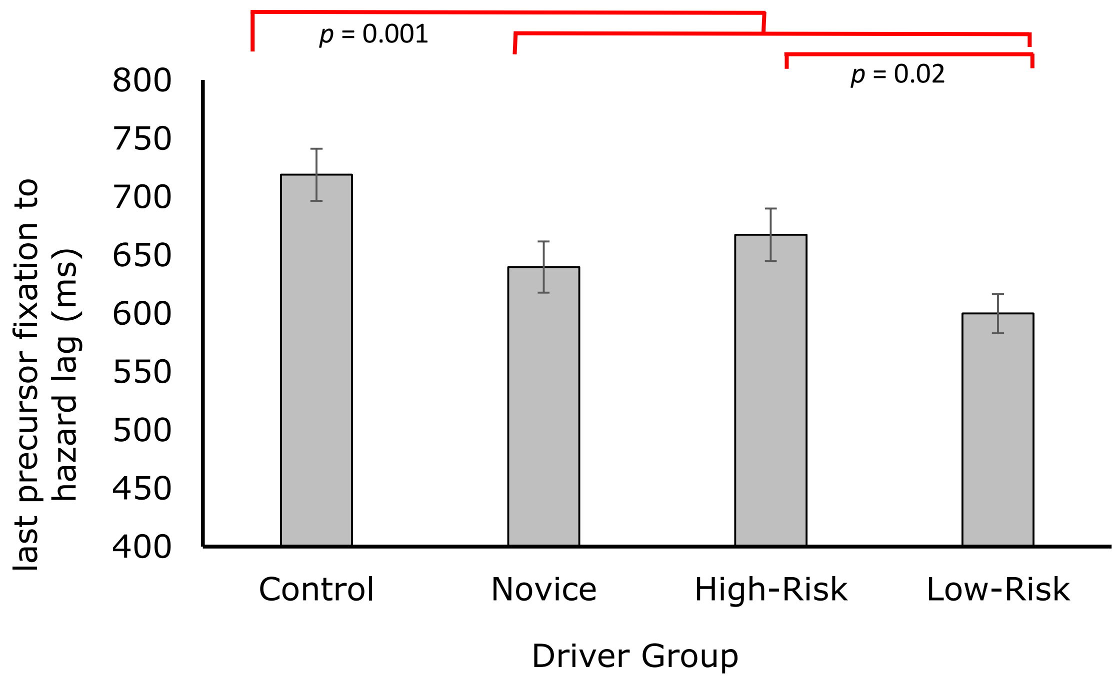


Figure 8. The average last time to fixate on the hazardous precursor for each Driver Group (with standard error bars added).

Several measures were recorded to reflect the amount of attention that participants gave to the hazard precursor. These included first fixation duration (the length of the first fixation given to a precursor by a participant), mean fixation duration (the average duration of all fixations given to each precursor), the number of fixations on each precursor, and the dwell time on precursors (the number of eye tracking samples that fell on the precursor during the prediction window). All of these measures were compared across the four driver groups, but only the analysis of the number of fixations proved to be significant, $F(13, 75) = 4.11$, $MSe = 0.01$, $p = 0.009$. Planned Helmert contrasts revealed that all fire-appliance drivers made significantly more fixations on the hazard precursors than the control participants (0.6 vs. 0.5; $p = 0.006$). Low-risk fire-appliance drivers also made significantly more fixations on the hazard precursors than the high-risk drivers (0.7 vs. 0.6; $p = 0.05$; see Figure 9). As all these means are lower than 1 fixation on the precursor, the data are very similar to those reported in Figure

7, though the addition of rare multiple fixations on the precursor pushes the difference between high and low-risk drivers over the significance threshold.

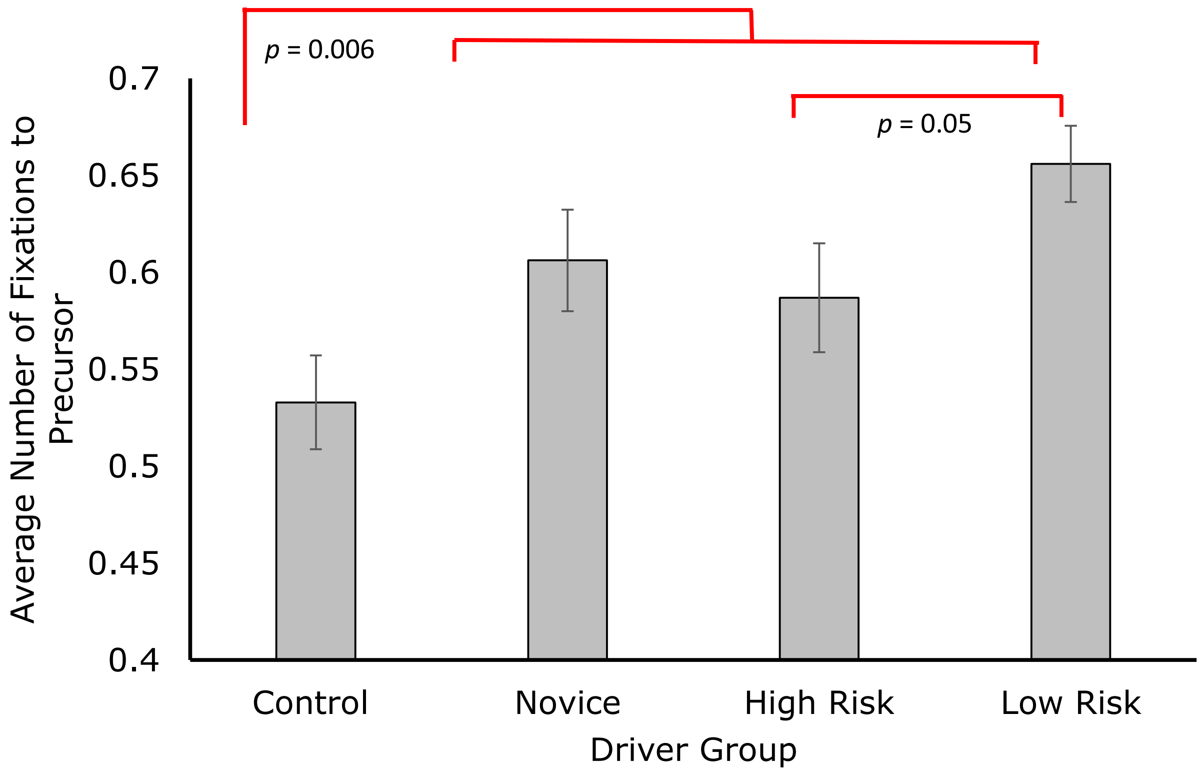


Figure 9. The average number of fixations on each hazardous precursor for each Driver Group (with standard error bars added). Note: these scores were converted back from Z-scores.

Discussion

The results of the hazard prediction test stand in contrast to those of the hazard perception test. The behavioural responses (RTs) to the hazard perception test (Experiment 1) only demonstrated a difference between the control group and the fire-appliance drivers taken as a whole. The behavioural responses to the prediction test (prediction accuracy) not also demonstrated a distinction between the control group and the fire-appliance drivers, but the low-risk group were also found to perform significantly better than the high-risk group. Thus the hazard prediction test has been more successful in discriminating between fire-appliance driver groups than the hazard perception test. This follows the pattern of results found by Ventsislavova et al. (submitted) albeit in a very different driving context. Ventsislavova et al found greater discrimination with a prediction test than a hazard perception test when comparing novice and experienced drivers from different countries. The current results demonstrate that the prediction test can be equally effective at discriminating on the basis of self-reported risk (rather than just experience) and can do so in a professional driver context that involves the highest levels of driver training.

The rationale behind the hazard prediction test is that safe drivers correctly prioritise and monitor potential precursors that may lead to hazards, and are therefore more likely to be looking in the right place at the right time. The current eye tracking results provide the first evidence in favour of this rationale, with the safest drivers being more likely to fixate the relevant precursor, and to be last looking at the precursor at the closest point in time to it becoming an actual hazard.

One alternative interpretation of these eye movement results is that the late fixations on precursors shown by the low-risk drivers might actually reflect the fact that they have only just looked at it. However, the groups do not significantly differ on how quickly they initially look at the precursors (and the means suggest a trend in favour of the safest drivers being the first to fixate the precursor, as well as being the last to fixate it). The low-risk drivers also make more fixations on the precursors than other drivers, though they do not differ in terms of overall dwell, suggesting that they may be monitoring other potential precursors with overt attention, returning to the precursor with the greatest evidence of becoming a hazard.

General Discussion

The aim of this study was to create a test that could discriminate between groups of safe and less-safe fire appliance drivers in order to better identify training needs. As a surrogate for safety, we categorised our drivers according to experience of driving fire appliances, and their self-reported safety (based on frequency, severity and responsibility for past collisions). The stimuli were designed to capture both the view from the specific vehicle and the visual demands of the actual task, and were thus filmed from fire appliance under realistic blue-light conditions (an approach used only once previously by Crundall et al., 2003, 2005, whose videos were appropriated from real dash-cam footage from police vehicles, but were of relatively poor visual quality).

Two variants of the hazard perception test were created: a traditional push-button hazard test requiring speeded responses to hazards, and a prediction test that provided participants with 4 possible outcomes for each clip following occlusion at the point of hazard onset. The hazard prediction test was the more successful of the two, successfully discriminating between the two highly-experienced groups of FA drivers, as well as differentiating all FA drivers from controls, on the basis of a percentage score for correctly predicted hazards (out of 15). The more traditional hazard perception test did not fare so well: the behavioural measure of response times could only discriminate between controls and all FA drivers. While this is in line with the literature which argues that emergency service staff have better hazard perception skills than control drivers in normal driving scenarios (Johnston & Scialfa, 2016; McKenna & Crick, 1991; Horswill et al., 2013), its lack of discrimination between the FA groups renders the perception test a poor potential tool for fire service instructors.

In addition to greater discrimination between groups, the prediction test also provides a simpler scoring methodology, readily understandable by future users. A score out of 15, or a percentage accuracy, is an unambiguous figure that demonstrates how well one performed in a test. Calculation of response times, however, raises many questions. The selection of the temporal scoring window is a particular concern, with internet forums full of complaints that those taking the UK test are penalised for pressing too soon (see Crundall, 2016). Even when the scoring window accepts a valid response, different research groups process the resultant response times in different ways. While many researchers might reference a favoured study whose methodology they follow (as we do with Wetton et al., 2010), there is no agreed method for dealing with missing values, skewed distributions, and non-standardised response windows. Some researchers have suggested novel approaches to dealing with these issues (e.g. survival analysis, Parmet, Meir and Borowsky, 2014), though by removing response times from the test completely we can avoid all such problems, while creating a more transparent scoring method for the average user.

It should be noted that in absolute terms, the significant differences between the driving groups are small. Are these still meaningful? The narrowness of these significant gaps between the high-risk and low-risk drivers reflects the fact that some high-risk drivers perform well on a prediction test, while some low-risk drivers still perform poorly. This is symptomatic of the fuzziness underlying the use of self-reported collision history to define our groups. Some drivers classed as low-risk might actually be quite dangerous on the road, but have still managed to avoid a serious collision, while other ‘low-risk’ drivers may have failed to report collisions in order to portray a safe image to researchers. Some drivers acknowledged they had been involved in other collisions that either were not worth rating (e.g. damage was inconsequential) or were too long ago to remember in detail, but it is possible that some of these collisions were more severe than participants admitted.

Conversely, some of our ‘high-risk’ drivers might be relatively safe. The collisions that led to their high-risk classification may have had mitigating circumstances that were not accounted for in our calculation, or their skills may have simply improved over time, possibly even as a direct result of a crash (e.g. Rajalin and Summala, 1997, found professional heavy-vehicle drivers were the only sub-group of their sample to demonstrate prolonged favourable changes in driving style following a fatal collision). Given the likely underlying fuzziness between our high and low-risk categories, a significant effect is all the more impressive. Also, were the current test to ever be used in a diagnostic capacity, one would not set the cut-off to catch all ‘high-risk’ drivers as defined in this study. Instead, only the extremely poor scorers would be targeted for further training.

One further problem with defining our risk groups is the question, what is it that makes them risky: errors of performance or volitional risk taking? The hazard prediction test is designed to detect problems in identifying upcoming hazards, but will not measure risk-taking behaviour. Looking at participant scores on the DBQ, it appears that our high-risk drivers suffer from both errors and slips/lapses more so than our low-risk drivers, yet they also score more highly on the violations factor. Thus our high-risk drivers represent a mixture of reasons that may account for their previous collision history, yet the hazard prediction test should only be discriminating these drivers from the low-risk group on the basis of errors. This further confusion of what constitutes a high-risk driver may have also weakened the effect. For future research it would be beneficial to separate out those drivers who are considered high-risk primarily due to errors from those who report high violation scores.

Conclusions

Both tests have demonstrated that fire appliance drivers have safer responses to filmed hazards compared to control, responding faster to hazards that appear, and predicting a greater number of correct hazards following occlusion. The hazard prediction test however has proved more effective in identifying differences between sub-groups of fire appliance drivers based on self-reported risk, and this is reflected in the eye movements of our drivers. The success of the prediction test over the hazard perception test is all the more impressive given that both tests used the same clips. This demonstrates that the occlusion methodology, with a purer measure of hazard prediction accuracy, is responsible for the improvement in discrimination rather than any differences across stimuli. The success of this test paves the way for a diagnostic test of hazard prediction for fire appliance drivers that will allow training resources to be better targeted, while the stimuli also offer new potential methods for training these skills in the future.

References

Becker, L. R., Zaloshnja, E., Levick, N., Li, G., & Miller, T. R. (2003). Relative risk of injury and death in ambulances and other emergency vehicles. *Accident Analysis and Prevention, 35, 6,* 941- 948.

Borowsky, A., Shinar, D., & Oron-Gilad, T. (2010). Age, skill, and hazard perception in driving. *Accident Analysis and 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., López Ramón, M.F., & Crundall, D., (2014). A first development and validation of the Spanish Hazard Perception Test. *Traffic Injury Prevention, 15, 8,* 817–826.

Chapman, P. R., & Underwood, G. (1998). Visual search of driving situations: Danger and experience. *Perception, 27, 8,* 951-964.

Chief Fire Officers Association (2015) Budget cuts will affect how fire service respond to emergencies [Press release]. Retrieved from http://www.cfoa.org.uk/news?tag=Press%20Release%202015.

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

Crundall, D., Chapman, P., France, E., Underwood, G., & Phelps, N. (2005). What attracts attention during police pursuit driving? *Applied Cognitive Psychology, 19, 4,* 409-420.

Crundall, D., Chapman, P., Phelps, N., & Underwood, G. (2003). Eye movements and hazard perception in police pursuit and emergency response driving. *Journal of Experimental Psychology: Applied, 9, 3,* 163.

Crundall, D., Underwood, G., & Chapman, P. (1999). Driving experience and the functional field of view. *Perception, 28, 9,* 1075-1087.

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

Deery, H.A., 1999. Hazard and risk perception among young novice drivers. *Journal of Safety Research 30, 4*, 225–236.

Drummond, A.E., (2000). Paradigm lost! Paradigm gained? An Australian's perspective on the novice driver problem. In: *Proceedings of the Novice Driver Conference*, Bristol, June 1–2.

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., & McKenna, F. P. (2004). Drivers' hazard perception ability: Situation awareness on the road. In: Banbury, S., Tremblay, S. (Eds.), *A Cognitive Approach To Situation Awareness*. Ashgate, Aldershot, UK, pp. 155–175. Horswill, M.S., Taylor, K., Newman, S., Wetton, M., Hill, A., 2013.

Horswill, M. S., Taylor, K., Newnam, S., Wetton, M., & Hill, A. (2013). Even highly experienced drivers benefit from a brief hazard perception training intervention. *Accident Analysis & Prevention, 52*, 100-110.

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.

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.

Johnston, K., A. & Scialfa, C., T. (2016). Hazard Perception in Emergency Medical Service Responders. *Accident Analysis and Prevention, 95*, 91-96.

Lim, P. C., Sheppard, E., & Crundall, D., (2014). A predictive hazard perception paradigm differentiates driving experience cross-culturally. *Transportation Research Part. F, 26,* 210-217

Lorains, M., Ball, K. & MacMahon, C., (2013). An Above Real Time Training Intervention for Sport Decision Making. *Journal of Psychology of Sport and Exercise, 14,* 670-674.

Maguire, B. J., Hunting, K. L., Smith, G. S., & Levick, N. R. (2002). Occupational fatalities in emergency medical services: A hidden crisis. *Annals of Emergency Medicine, 40, 6*, 625-632.

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

Maycock, G., Lockwood, C. R. & Lester, J. (1991). The accident liability of car drivers. TRRL Report RR315. Crowthorne: Transport and Road Research Laboratory.

McKenna, F. P., Horswill, M. S., & Alexander, J. L. (2006). Does anticipation training affect drivers' risk taking? *Journal of Experimental Psychology: Applied, 12, 1,* 1-10.

McKenna, F. P., & Crick, J. L. (1991). Hazard perception in drivers: a methodology for testing and training. Crowthorne, UK: Transport and Road Research Laboratory.

Özkan T, & Lajunen T. (2005). Multidimensional Traffic Locus of Control Scale (T-LOC): factor structure and relationship to risky driving. *Personality and Individual Differences*, *38*, 533-545.

Parker, D., Reason, J.T., Manstead, A.S.R., & Stradling, S.G., (1995). Driving errors, driving violations and accident involvement. *Ergonomics, 38,* 1036–1048.

Parmet, Y., Meir, A., & Borowsky, A. (2014). Survival analysis: A fresh approach for analyzing response times in driving-related hazard perception paradigms. *Transportation Research Part F, 25,* 98-107.

Pelz, D.C., & Krupat, E., (1974). Caution profile and driving record of undergraduate males. *Accident Analysis and Prevention, 6,* 45–58.

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

Pradhan, A. K., Pollatsek, A., Knodler, M., & Fisher, D.L (2009). Can younger drivers be trained to scan for information that will reduce their risk in roadway traffic scenarios that are hard to identify as hazardous? *Ergonomics, 52, 6*, 657-673.

Rajalin, S., & Summala, H., (1997). What surviving drivers learn from a fatal road accident. *Accident Analysis and Prevention, 29, 3,* 277–283.

Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K., (1990). Errors and violations on the road: a real distinction? *Ergonomics*, *33*, 1315–1332.

Sagberg, F., & Bjørnskau, T. (2006). Hazard perception and driving experience among novice drivers. *Accident Analysis and Prevention, 38, 2,* 407–414.

Scialfa, C. T., Deschênes, M. C., Ference, J., Boone, J., Horswill, M. S., & Wetton, M. (2011). A hazard perception test for novice drivers. Accident Analysis & Prevention, 43(1), 204-208.

Underwood, G., (2007). Visual attention and the transition from novice to advanced driver. *Ergonomics, 50, 8,* 1235–1249.

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

Watts, G.R. & Quimby, A.R. (1979). Design and validation of a driving simulator for use in perceptual studies. Report 907. Crowthorne, UK: Transport Research Laboratory.

Wells, P., Tong, S., Sexton, B., Grayson, G., & Jones, E. (2008). *Cohort II: A study of learner and new drivers. volume 1*. Crowthorne, Transport Research Laboratory.

Wetton, M. A., Horswill, M. S., Hatherly, C., Wood, J. M., Pachana, N. A., & Anstey, K. J. (2010). The development and validation of two complementary measures of drivers' hazard perception ability. *Accident Analysis & Prevention*, *42, 4,* 1232-1239.

Ventsislavova, P., Crundall, D., Castro, C., Gugliotta, A., Garcia-Fernandez, P., Zhang, W., Ba, Y., & Li, Q., (submitted). A cross-cultural comparison of hazard perception and hazard prediction. *Accident Analysis and Prevention.*

Ventsislavova, P., Gugliotta, A., Peña-Suarez, E., Garcia-Fernandez, P., Eisman, E., Crundall, D, & Castro, C. (2016). What happens when drivers face hazards on the road? *Accident Analysis and Prevention, 91*, 43-54

Zuckerman, M., Kolin, E. A., Price, L., & Zoob, I., (1964). Development of a sensation-seeking scale. Journal of consulting psychology, 28, 6, 477.

Footnotes

1. A fire appliance is large liveried vehicle, mounted with sirens and flashing lights, which is designed to transport a variety of rescue equipment, and fire-fighting media (e.g. water, foam). It has a raised driving position, and can usually carry 6 fire-fighters in the cabin. Fire appliances are also called fire engines, fire trucks and fire tenders.
2. Malone and Brünken (2015) have also compared multiple-choice questions to response times, but their questions appeared after the hazards had been passed by the film car, and were therefore not designed to capture online measures of hazard prediction. The authors referred to their multiple-choice trials as having low ecological validity.