Creating a hazard-based training and assessment tool for emergency response drivers

Victoria Kroll, Andrew K. Mackenzie, Thomas Goodge, Rowena Hill, Robert Davies, and David Crundall*

Nottingham Trent University

*Address for Correspondence:

Professor David Crundall

Department of Psychology

Nottingham Trent University

50 Shakespeare Street.

Nottingham

NG1 4BU.

Keywords: hazard perception, hazard prediction, What Happens Next? emergency response driving.

Abstract

Emergency response drivers (ERDs) are often required to engage in high-risk driving manoeuvres on their way to a reported incident. Such risk requires that these drivers receive a high-level of training and continued development. The aim of this paper was to investigate an innovative format for a new potential tool that could support the training and assessment of these drivers: a single-clip Holistic Hazard Test, containing multiple hazards in a single route. In study one, we created a proof-of-concept 15-minute clip containing hazards, multiple-choice questions and probes to collect self-reported safety ratings. ERDs were more accurate on the multiple-choice questions (MCQs) than a control group, though response time scores to hazards did not reach the threshold for significance. In study two, we refined the development process and created a series of new holistic hazard tests across four counties of the East Midlands, UK. Each test contained many hazards and MCQs that assessed situation awareness and decision-making, based on the results of study 1. Participants were recruited across the four counties and were presented with both the test that was specific to their county and one of the unfamiliar-location tests, in order to assess the generalisability of the tests across different locales. The results showed no differences regarding location familiarity, suggesting that tests filmed in one area of the country can be viewed by drivers elsewhere without detriment to performance. ERDs once again responded to MCQs more accurately, and also scored more hazard points on the basis of faster responses to hazards compared to control participants. These results suggest such tests can successfully tap into ERD-specific skills with regard to spotting, predicting and responding to hazards on the road. We recommend refinement of this tool for assessment of emergency response drivers, and further development to extend the materials to create a training tool.

Introduction

Emergency response drivers (ERDs) are frequently required to engage in high-risk driving manoeuvres in an effort to arrive at a reported incident as fast as possible, while ensuring the safety of all road users *en route*. They often drive at a speed greater than that of the prevailing traffic, and may contravene road rules to progress, such as driving through a red traffic light, or passing on the wrong side of a central lane divider. Despite advanced training, the nature of the job places these drivers in a high-risk category for collisions (e.g. Becker et al., 2003; Crundall et al., 2003). One study, for instance, found that medical, police, and fire and rescue personnel had a significantly greater occupational fatality rate than the general public, with traffic collisions playing a major part in this discrepancy (Maguire et al., 2002). Such levels of on-road risk necessitate a high level of initial driver training and subsequent skill maintenance, though any further additions to already intensive training programs are currently hampered in the UK by substantial budget cuts. For example, in the UK fire service alone, budgets were recently targeted to be cut by 22%, continuing a decline in funding that began in 2010 (Chief Fire Officers Association, 2015). This combination of a high-risk task with increasingly stringent budget constraints requires the emergency services to seek innovative and cost-effective methods to supplement current on-road training.

One potential method is to include Hazard Perception (HP) testing as part of emergency service driver training and assessment. Hazard perception skill reflects a collection of sub-processes that include visual search for hazardous precursors, prediction and prioritisation of potential hazards, and then detection and processing of a hazard if it occurs (Pradhan & Crundall, 2017). The ability to detect driving hazards has been the focus of a substantial amount of research using a wide variety of methodologies and stimuli, though the common form of the modern test requires participants to watch video clips taken from the driver's perspective and make a response, usually a button press, when they detect a hazard.

Over 50 years of research in this area has demonstrated that video-based hazard perception tests can differentiate between self-reported collision-free and collision-involved drivers (e.g. Pelz & Krupat, 1974; Watts & Quimby, 1979; McKenna and Crick, 1991), while participant scores can even predict likelihood of being involved in a future crash (Drummond, 2000; Boufous et al., 2011). Hazard perception tests have also been found to be sensitive to levels of driver training and experience, with expert or experienced drivers often out-performing less-experienced drivers (Renge, 1998; Wallis &

Horswill, 2007; Horswill et al., 2008; Deery, 2000; Pradhan et al., 2009). Taken together, these findings suggest that a high level of HP skill is related to a reduced likelihood of having a crash, and that this skill is developed with experience over time, or through advanced training. Indeed, the lack of hazard perception skill in novice drivers has been blamed, at least in part, for the over-representation of young, novice drivers in the crash statistics both in the UK and in many other countries (Crundall et al., 2012; Braitman et al., 2008; Maycock et al., 1991; Underwood, 2007). Interestingly, however, it appears that HP skill does not reach an obvious ceiling, with even highly experienced drivers benefiting from HP training (Horswill, Taylor, Newnam, Wetton, & Hill 2013).

Based on such evidence, the U.K. government introduced a national HP test to the U.K. driver-licensing procedure in 2002. The rational for such a test is grounded in the assumption that those learner drivers who spot and respond to hazards more quickly will be more likely to avoid similar hazards in the real world, decreasing their probability of crashing (e.g., McKenna & Crick, 1991; McKenna & Horswill, 1999; Quimby et al., 1986). Research following the introduction of the HP test suggested that the introduction of the test led to an 11.3% decrease in on-road collisions that did not involve low-speed manoeuvres (Wells et al., 2008; though the lowest level of the 95% confidence interval was much less at 0.3%), presumably either by keeping the worst drivers off the roads, or by ensuring that learner drivers are trained in hazard perception by their instructors in preparation for the test. Regardless of the underlying mechanism, the introduction of the HP test is considered to have been successful in reducing on-road collisions. One researcher has equated this to a *per* year reduction of 8,535 damage-only collisions, and 1,076 fewer injury collisions, with an estimated annual saving to the UK of nearly £90 million (Horswill, 2016).

While there are many decades of research supporting the efficacy of HP testing under normal driving conditions, can this methodology be applied to emergency response driving? A few published studies have tested ambulance and police response drivers (Johnston & Scialfa, 2016; McKenna & Crick, 1991; Horswill et al., 2013) and have found them to out-perform control drivers on speeded responses to videoed hazards. This difference may be due to the advanced training received by emergency service drivers and their experience gained in highly hazardous situations, though self-selection for high-risk driving jobs may also contribute to the effect.

In the above-mentioned studies, the HP tests were comprised of clips of civilian driving. If one's aim is to assess whether advanced levels of training and driving experience will improve hazard perception performance during normal driving, then this approach is perfectly suited to the task. However, if we want to assess emergency response drivers' ability to spot hazards while driving under response conditions, then the test needs to reflect that driving context. Response drivers will typically engage

in manoeuvres that contravene road rules, placing themselves in situations where the nature and magnitude of subsequent hazards are likely to differ from ordinary driving. In addition, the approach of a liveried vehicle with flashing lights and sirens can provoke responses in other drivers that are much less predictable than typical road-user behaviour.

Unfortunately, the pragmatics of recording footage from real emergency-service driving pose a barrier to the development of an emergency-response hazard perception test. To date there are only a limited number of studies that have examined the differences between emergency response drivers and ordinary drivers using clips filmed in real emergency response situations (Crundall et al., 2003, 2005).

Crundall et al., (2003; 2005) found that although police drivers did not overtly identify more hazards, they did have a greater horizontal spread of search (consistent with a more efficient search strategy; Underwood et al., 2002), spent more time looking at sources of potential hazards (e.g., pedestrians, parked vehicles and side roads), and produced a greater number of electrodermal responses than the controls. This suggests that the police drivers were potentially aware of a greater number of arousing stimuli than control drivers, though this did not translate into a behavioural response. One possible explanation is criterion bias: police drivers may have spotted hazards early but did not press immediately because they thought the threat posed was within their skill levels to avoid.

In a more recent study (Crundall and Kroll, 2018), video footage was taken from fire-appliances on blue-light training runs and three groups of fire-appliance drivers were tested (novice, low-risk and high-risk), along with age/experienced-matched control drivers. They found fire-appliance drivers responded faster to hazards than the control drivers, though no differences were observed between the three fire-appliance driver groups. Interestingly, eye-movement analyses suggested that the experienced fire-appliance drivers were faster to fixate the hazards, looked at more hazards, and spent more time looking at the hazards than the novice fire-appliance drivers. This is in line with previous research that has observed eye-movement differences between groups of drivers, which have then failed to translate into subsequent response-time differences (e.g., Chapman & Underwood, 1998; Crundall et al., 1999). Although the stimuli can elicit subtle behavioural differences in eye movements, the primary response-time measure of the hazard perception test appears too insensitive to detect these differences.

As part of the same study, Crundall and Kroll (2018) also created a *hazard prediction* test based on the fire-appliance clips, which proved to be more successful in identifying response differences between the three groups of fire service drivers. This type of test differs to the more-traditional hazard perception test in that each clip occludes at hazard onset, and participants are asked “What happens

next?" (Jackson et al., 2009, Crundall, 2016, Ventsislavova et al., 2019). Participants are then provided with four possible outcomes to choose between. If they were looking in the right place at the time, immediately prior to occlusion, participants can easily identify the correct answer. To ensure that they are looking in the right place at the right time however, participants need to engage in a constant process of prioritising and predicting potential hazards.

This modified hazard test found low-risk fire-appliance drivers to be more accurate at predicting the hazard than high-risk fire-appliance drivers. The hazard prediction test was originally proposed as an alternative to the hazard perception test, with the aim of mitigating some of the problems with traditional hazard perception tests (e.g. criterion bias, subjective scoring windows, see Crundall et al., 2016). Subsequent evidence has demonstrated the robustness of this test in differentiating between safe and less safe drivers based upon the assumption that safer drivers are better able to prioritise hazardous precursors and therefore better able the predict the hazard (Jackson et al., 2009; Castro et al., 2014; Crundall, 2016; Lim et al., 2014; Ventsislavova et al., 2016, 2019). Taken together, these findings suggest that low-risk fire-appliance drivers were better able to monitor and correctly prioritise hazardous precursors. The hazard prediction test also has the benefit of ease of scoring as each participant gets a percentage accuracy score that is easily calculated.

The current study reflects only the third attempt in the literature to test emergency response drivers' hazard perception skills using footage captured from realistic *blue-light* runs (i.e. response-drives involving contravention of road rules where necessary, and the use of blue flashing lights and sirens). It was undertaken to assess the possibilities of designing a future method of assessing (and training) hazard-relevant skills in the emergency services, as a cost-effective supplement to existing on-road training provision. The stimuli were captured from blue-light training runs using an Emergency Light Vehicle (ELV) travelling on urban, suburban and arterial roads, providing ample opportunity to capture footage of hazardous situations. While Crundall et al. (2003; 2005) had to design their study around a limited batch of previously-collected clips, the current study had the opportunity to build a new test from the ground-up, allowing us the freedom to design something much more tailored to our current understanding of the hazard perception process.

The sub-processes of hazard avoidance

Pradhan and Crundall (2017) chose the term Hazard Avoidance to describe the whole process of safely navigating past a hazard, which includes searching for hazardous precursors, prioritising precursors for subsequent monitoring, predicting upcoming hazards on the basis of their precursors, mitigating the potential for a hazard to occur, processing and appraising hazards if they do occur, and finally

selecting a suitable response. Selection of a response includes choosing both the actual action (e.g. braking, swerving in a particular direction) and the magnitude of the action (e.g. gradual or harsh).

Hazard perception response times typically confound many of these sub-processes. Slower response times can be due to poor hazard searching, a failure to identify precursors, a failure to predict the hazard, slow processing of the hazard, or inappropriate appraisal of the perceived risk relative to perception of one's own skill. There have, however, been studies that have attempted to measure some of these sub-processes in isolation. For instance, experienced drivers have been found to fixate hazardous precursors sooner and more frequently than inexperienced drivers (Borowsky et al., 2010; Crundall et al., 2012; Pradhan et al., 2005).

Appraisal of risk in HP scenarios has also been investigated, typically employing rating scales that are presented after each video clip has ended. Across several studies it has been noted that young and inexperienced drivers will rate situations as less risky or hazardous than their more experienced counterparts (Matthews & Moran, 1986; Renge, 1998; Scialfa et al., 2012), though other studies have failed to find this effect (Crundall et al., 1999; Wallis & Horswill 2007).

Selection of an appropriate manoeuvre following the detection of a hazard is typically investigated in simulators, where a naturalistic response can be recorded. For instance, Hancock and Ridder (2003) found that swerving was favoured over braking if the available time to view the hazard was extremely short (e.g. approaching a head-on collision with another vehicle obscured by the brow of a hill), though braking and gentle deceleration was preferred on hazards with a relatively long lead-in time. But does hazard response vary with experience? Certainly, the magnitude of the response tends to vary systematically with experience and training, with inexperienced or untrained drivers showing the smallest reductions in speed in the presence of hazards in simulated environments (e.g. Mueller and Trick, 2012; Crundall et al., 2010), but this is very difficult to measure in a video-based hazard perception test. Selection of an appropriate response is however more amenable to testing via an HP methodology. For instance, Malaterre et al. (1988) predated Hancock and Ridder's (2003) study by 15 years, using a combination of video clips and slides to judge drivers' response selection when faced with a hazard. Their participants reported favouring a steering response rather than a braking response when hazards appeared close to their 'vehicle'. Despite the potential for measuring response selection in hazard perception tasks, few studies have compared response selection across safe and less-safe drivers. One exception was published by Ventsislavova et al. (2016). They asked drivers to decide whether they would make an evasive manoeuvre or not following the presentation of hazardous and quasi-hazardous video clips. Inexperienced drivers were less likely to report the need for an evasive manoeuvre than more experienced drivers, though as they were also less sensitive to

the appearance of hazards (d') it is difficult to claim that their manoeuvre decision was independent of whether they spotted the hazard or not.

The above studies offer numerous ways to address the sub-components of the whole hazard avoidance process (that is, all the sub-components that comprise the typical definition of hazard perception, plus mitigation behaviours and the selection of an appropriate response). Within the current paper, we questioned whether it would be possible to create a hazard perception test that benefits from several different measures of hazard perception skill (or more inclusively, *hazard avoidance* skill). Thus, in addition to capturing hazard perception stimuli that are specific to emergency service drivers, we created a test that combined the typical speeded-button response to developing hazards, with a series of multiple-choice questions (MCQs) that appear at pauses within the video footage to probe various sub-components of hazard avoidance skill. We have called this test the *Holistic Hazard Avoidance test* to reflect the fact that it probes a wide range of hazard-related sub-processes, including prediction ("What happens next?") and response selection ("What would you do next?"), which fall outside the typical definition of a *hazard perception* test.

The current studies

Our Holistic Hazard Test (HHT) combines a range of measures to probe hazard awareness, using naturalistic footage captured from an emergency light vehicle (ELV) across several counties in the East Midlands on blue-light training runs. Here we present two studies that attempt to assess the use of such emergency-service HHTs.

In study one, a 15-minute clip was selected from 10 hours of footage from Nottinghamshire containing 13 identified hazards. Hazards were caused by the actions of other road users who were either oblivious to the approach of the ELV, or who made poor response selections while trying to get out of the way of the emergency vehicle. The measures taken included traditional speeded-responses to hazards while the clip was playing, answers to multiple-choice questions provided within pauses during the video, and safety ratings on a Likert scale also presented during pauses in the playback. The multiple-choice questions probed a number of hazard avoidance sub-processes including comprehension of the roadway, prioritisation of potentially hazardous locations, prediction of hazardous events, and response selection. A group of trained ERDs were tested on the HHT. Their behavioural responses to the hazards (response times), safety ratings, and answers to the multiple-choice questions were compared to the performance of a control group of drivers. We predicted that this novel and potentially encompassing test of hazard avoidance would discriminate between these two groups.

In study two, we refined the test-development process and created four new tests for different counties of the East Midlands in the UK (Nottinghamshire, Lincolnshire, Derbyshire and Leicestershire). A larger group of emergency service drivers were recruited to take part, allowing us to compare both high and low-experienced response drivers with a control group. Drivers were recruited from across the four counties and took part in a 'local' test, and a 'non-local' test. This allowed us to assess the impact of geographical familiarity on the ability of the tests to differentiate between driver groups. This manipulation was designed to answer the question of whether each emergency service requires its own locally-sourced hazard test, or whether 'one size fits all'.

Study One

Method

Participants

Twenty emergency response drivers and twenty control drivers were recruited for the study. The emergency light vehicle drivers were recruited from the fire service and were either command car drivers (15) or first responders (5), both having received emergency light vehicle training. These drivers were all male, with a mean age of 45 years, 15,150 personal miles per annum, and 25 years of driving experience since passing their driving test. They had driven fire service vehicles for an average of 13 years (range: 1.5 months to 25 years of experience) and reported undertaking blue-light runs in a car an average of 7 times per month. Control drivers included 8 males, with a group mean age of 31 years, 5750 personal miles per annum, and 11 years driving experience since passing the test.

Stimuli and Apparatus

Filming

The test was developed from footage that was captured from a single emergency light vehicle (ELV) on an emergency response training run. The clip was filmed around Nottinghamshire in April 2015. The filming took place during driver training to avoid the necessity of undertaking additional non-emergency blue-light runs beyond those required for training purposes. In total, approximately 10 hours of footage was obtained from the ELV.

Filming from the car required a 4-camera system in order to capture the forward view from the cabin and the 3 views that are available to the driver through the mirrors. A GoPro HERO4 Silver Edition camcorder recording in Full High Definition (1080p, 16:9 ratio, wide-angle setting) was mounted on a suction cup to the interior of the windscreen to capture the forward view. Two JVC Action Cameras (Model Number: GC-XA1BU; 1080p, 16:9 ratio) were mounted externally using suction mounts aligned

with the side mirrors but positioned to avoid obstruction for the driver. These cameras were tethered to the car for safety. One further JVC cameras was mounted on the inside of the rear windscreen to capture the rear-view mirror information. The recorded mirror views, and the forward-facing view, were subsequently combined with a graphic overlay of a car interior to create an immersive experience (see Figure 1. for a screen shot from the finished clip).

All hazardous events arose from opportunistic on-road filming whilst the vehicles were progressing under blue lights with periodic sirens.

Figure 1. A screen shot of the final edited clip used for the ERDs Holistic Hazard Avoidance test showing the combined video feeds from the forward-facing camera and three rear-facing cameras (which were edited into the mirror locations depicted in the graphic overlay of the car interior). This screen shot is taken from a point where the video has paused, and a question is presented on the screen. Participants must respond to this question by selecting the correct answer via a computer keyboard in order to continue watching the video.

Editing and Clip Selection

Prior to video editing, a graphic overlay was designed to represent the interior of an ELV (see Figure 1.). A-pillars and the internal roof of the vehicle were 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 four cameras were then synchronised and edited in Adobe Premiere CC such that a passing vehicle would disappear from the forward view and reappear in the mirror views following a brief gap to represent the other vehicle passing through the blind spot of the film car. The footage

from the mirror cameras was positioned within the mirror placeholders available in the graphic overlay. As with typical hazard perception tests, the final clip did not have an audio track.

For the test a 15-minute clip was required that contained sufficient naturally-occurring hazards to encourage constant vigilance in the participant, and a number of *critical decision points* where probe questions could be inserted into a pause in the clip. We opted for a single clip with multiple hazards (rather than the more typical hazard perception format with multiple short clips) in order to increase immersion and provide the feeling of a complete emergency drive. In discussion with Fire Service Driving Instructors a suitable clip was identified. Two instructors viewed the clip independently and provided a think-aloud commentary. The commentaries of both instructors were compared, with overlaps in the commentaries providing the critical decision points, which would be subsequently probed by multiple-choice questions and rating scales. Thirteen *a priori* hazards were identified that would require speeded-button responses from participants. Twenty-one critical decision points were also identified in the 15-minute clip. At each of these points, the video was edited to pause and display the question.

The 21 questions fell into 5 categories: *safest way to proceed? (response selection), safety ratings, what happens next? (Prediction), roadway comprehension, and prioritisation of potential hazards.* None of these 21 questions referred to the 13 hazards that were identified for the speeded-button response part of the test.

Five questions examined the safest way to proceed, offering participants 4 options of how to proceed around a particular hazardous obstacle (e.g., 'What is the safest way to proceed? 1. Proceed around the nearside of the silver car; 2. Turn sirens off and wait behind the silver car; 3. Wait at a distance until the silver car moves; 4. Proceed towards the silver car until it moves'). During question presentation the paused video was visible.

Five questions assessed participants' ratings of how safe a specific road situation was. Participants gave answers on a 7-point Likert scale (ranging from 1 –'not at all safe' to 7 – 'completely safe'). For example, 'how safe is it to overtake the van?'. For this particular overtaking question, the views in the three mirrors were masked while the rating scale was presented, thus ensuring that participants could only use mirror information prior to the question, to inform their decision. For all other rating questions, the screen simply remained frozen during presentation of the rating scale.

Two items probed the drivers' ability to predict an imminent hazard (e.g., 'what happens next? 1. The silver car pulls out of the side street; 2. The oncoming bus strays into your lane; 3. A motorcyclist overtakes you; 4. Heavy congestion appears in your lane ahead'). These two items probed different

hazards to the 13 *a priori* hazards identified for the speeded-button responses. At the point of presenting the question the driving scene was occluded following Jackson et al., (2009) and Crundall (2016).

Four items probed drivers' roadway comprehension in terms of understanding of the current environment (e.g. 'which of these is true? 1. The left-hand lane closes in 300 yards; 2. There is an obstruction in 300 yards, 3. There are road works in 300 yards; 4. There are temporary traffic lights in 300 yards'). Typically, items referred to road signs that had just been passed. These road signs were not visible on the screen during the pause where the image remained frozen for the presentation of the question.

Finally, five items examined participants' abilities to prioritise potential hazard precursors. These questions asked drivers to identify the most likely source of a hazard within the scene at any one moment (e.g., 'What potential hazard should be taken into consideration here? 1. Pedestrians on the left; 2. Cars emerging from the junction on the left; 3. Pedestrians walking from in front of the van to the right; 4. Cars entering the junction from the left'). The image remained frozen and visible during the presentation of the question.

The test was presented to participants on a Lenovo X240 laptop with a 12.5" screen and built in keyboard. A separate Lenovo ball mouse was also used to record button presses to the hazards.

Design

A between-subjects design comparing driver experience (ERDs vs. controls) was used. The main dependent variable was the participants' response time (RT) to detect the hazards in seconds. RTs were calculated from the hazard onset time, that were chosen in discussion with emergency response driver instructors. Other dependent variables included the participants' accuracy at answering the 16 multiple-choice questions and the Likert ratings for the remaining 5 questions. The correct MCQ answer was based on answers given by the driving instructors. Correct responses were awarded one point each.

Procedure

Fire Service personnel were tested in a quiet office in their respective fire stations while on shift. Control participants were tested within a laboratory at the Nottingham Trent University. Following a demographic questionnaire, participants were seated approximately 60cm from the screen. They were informed that the footage they were about to watch was taken from the perspective of an emergency response driver, driving in an emergency response situation (i.e., a blue-light run).

Participants were instructed to watch the clip with two goals in mind. Their first task was to respond to potential hazards during the playback of the clip, by pressing a mouse button to indicate the presence of a hazard that would require them to suddenly stop, slow down or change position to avoid a potential collision. The second task was to respond to questions that would suddenly appear on the screen during a pause in the video playback. They were informed that sometimes the video would remain paused with a frozen image visible on the screen, while at other times the driving scene would be occluded during the presentation of the question. Participants were told that these questions would relate to their driving judgements, and that they would be required to press the appropriate number on the keyboard to record their response. When the questions were answered via the selection of a multiple-choice answer or by providing a rating, the playback of the footage was resumed. No feedback was given to participants regarding the accuracy of their responses during the test.

Results

Analyses were conducted to determine whether driving experience and training influenced participants' response times to hazards, their rating of safety (5 items) and their response accuracy to the 16 MCQs embedded in the video at the critical decision points.

Speeded-responses to the hazards

Across 20 ERDs and 20 control drivers, 12 out of the 13 key hazards received a response within the hazard window from over 50% of all participants (with 65.8% of participants responding to these hazards on average). One hazard however only received responses from 17.5% of all participants (3 control drivers and 5 ERDs). This hazard was removed from all further analyses as an outlier. The percentage of successful responses to the 12 remaining hazards were 66.7% and 64.1%, for EDRs and control drivers, respectively ($t(39) = -.58$, $p = 0.57$). Response rates to these 12 hazards had a Cronbach's alpha of 0.5.

Following the method used by the DVSA to score the national hazard perception test, the scoring windows for each hazard were split into 5 even sections, with 5 points awarded for a response in the first section, 4 points for a response in the second section, and so on. The resultant data were then compared via an independent t-test, which revealed a marginal effect of driver group, with the ERDs responding faster to the hazards on average compared to the control drivers (26.2 points versus 20.9 points; $t(38)= 1.81$, $p = 0.07$, *Cohen's d* = 0.56; see Figure 2a).

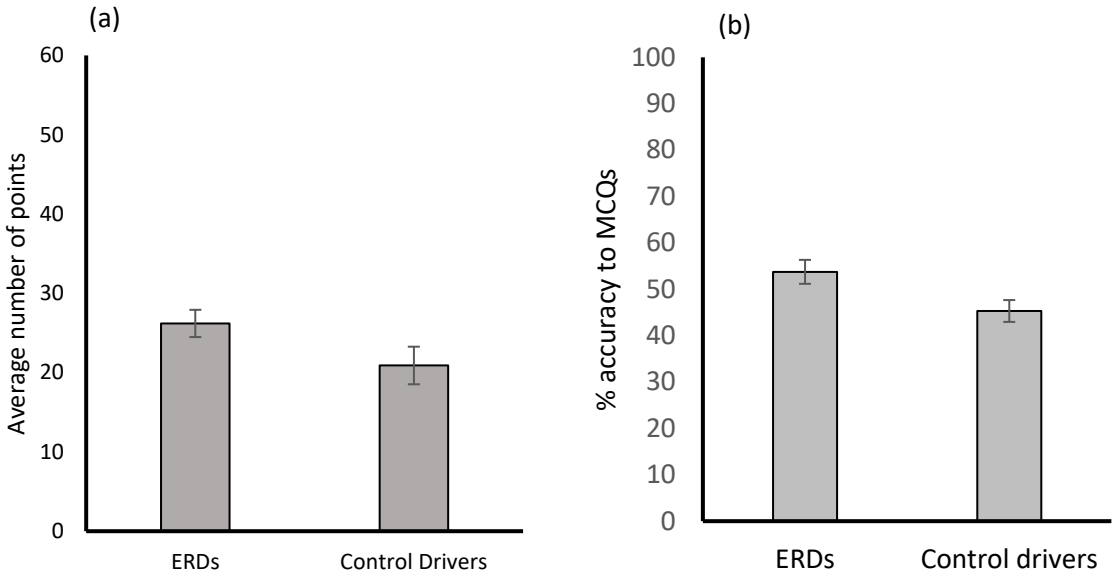

Figure 2. Panel A represents the average number of points for each driver group. Standard error bars are added. Panel B displays the mean percentage accuracy of responses to the 16 MCQs. Standard error bars are added.

Responses to the Multiple-Choice Questions

In addition to the five rating questions, participants were presented with 16 MCQs during the video clip. Percentage accuracies across the 16 questions were calculated for all participants and were subjected to an independent t-test comparing ERD performance to control driver performance. A significant effect was found ($t(38) = 2.42$, $p = 0.02$, $d = 0.8$) with ERDs outperforming the control drivers (53.8% vs. 45.3% respectively; see Figure 2b.).

To assess the contribution of individual questions to this effect, the percentage of ERDs and control drivers who responded correctly to each question was also charted (see Figure 3). This graph indicates that not all items contributed towards the overall superiority for ERDs on the MCQs. While many items suggest greater accuracy for ERDs over control participants, there are others where no difference is apparent, or even where the ERD advantage is ostensibly reversed. Interestingly, there does not appear to be a clear pattern between the four question types. Each category of question has items that seem to discriminate in favour of the ERDs. This suggests that all categories are potentially of use in creating valid items to discriminate between driver groups, though the detail in the individual items is likely to determine their ultimate validity. The Kuder–Richardson Formula 20 (KR-20) for each of the four MCQ question types were calculated, though unsurprisingly the resultant statistics were not high (primarily because we have chosen these items to be heterogenous). The KR20 for the prediction questions was the highest (0.34), though still some distance from a good level of internal consistency (0.7). Such lack of internal consistency is not uncommon with hazard perception tests (e.g. Horswill & Mckenna, 2004).

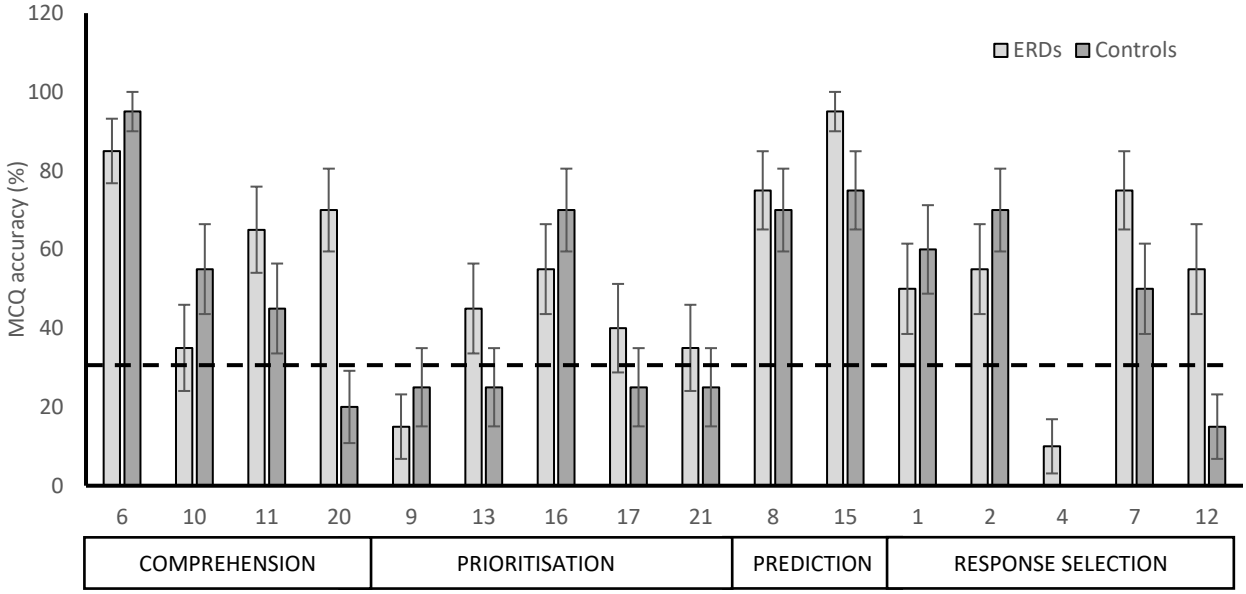

Figure 3. The percentage of participants who responded correctly to each of the 16 multiple-choice questions (MCQs). Standard error bars are added. The dashed line reflects the mean chance expectancy.

Responses to ratings

Of the 21 questions edited into the clip, five of these asked participants to rate how safe they thought a particular situation or manoeuvre was on a scale from 1 – 7 (where 1 is 'not safe at all' and 7 is 'completely safe'). The rating questions were Q3, Q5, Q14, Q18 and Q19 in the order of presentation. An average of all 5 ratings failed to discriminate between the two groups (t(38)) = -1.45, p = 0.157). Group comparisons at item-level did not reveal any differences.

Discussion

The primary aim of study 1 was to create a proof-of-concept holistic hazard avoidance test (HHA test) for emergency response drivers, which could then be refined in study 2. Two innovations distinguish this from traditional hazard perception tests. First, the stimuli were captured from real blue-light runs, providing a test that is specific to the emergency services. Secondly, this test included a variety of measures in contrast to the traditional format of other such tests. These measures probed a range of hypothesised sub-processes that comprise hazard avoidance skill. To identify whether the resultant test was sensitive to the training and experience of emergency service personnel, we compared the performance of a group of ERDs to that of a group of control participants across the key dependent variables.

The results identified that the MCQs revealed a significant group difference in performance, suggesting that the novel insertion of probe questions into a hazard perception test may be useful in future ERD assessment and training tools. With so few instances of each category of MCQ it is inappropriate to formally compare performance across these groupings. Yet the descriptive data (Figure 3) suggests that discriminative items may occur in all categories. Equally however, those items that appear to evoke identical performance from the two groups appear evenly distributed across the categories also. This suggests that no one category is responsible for the effect, but rather the overall significance is dependent on the specific details of each item. Furthermore, it is also apparent that several items do not appear to evoke average scores above chance (25%; see Figure 3). It is important to seek to overcome these issues with iteration, with successful items retained for future tests, while non-discriminative items are modified and reassessed.

The analysis of HP score, derived from speeded-responses to hazards, did not reach significance (p = .07), although the ostensible trend was in the predicted direction. The relatively low number of hazards responded to (even by ERDs) suggests that some of our *a priori* events may not have been considered hazardous by the drivers (Crundall et al., 2003). In a typical hazard test, we would simply replace the poorly performing hazards with new ones. However, in the current test we used a single clip to better provide the feeling of a single emergency response drive. This makes item replacement very difficult and suggests more care is needed in selecting the original clip to ensure that hazards are sufficient to evoke responses in the majority of cases.

One limitation of the current study is that the control drivers in study 1 were relatively less experienced than the ERDs (11 years versus 25 years, respectively). Given that hazard perception skill is gained via driving experience (Horswill, 2016), this experiential gap may have contributed to the significant group different in MCQ score and the marginal difference in HP score. Even though the control group would be considered highly experienced even by the most conservative estimates of highly-experienced drivers (e.g., Horswill & Mckenna, 2004), we acknowledge that the greater ordinary driving experience of the ERDs could have contributed to the effects. Alternatively, the younger controls could have had an unfair advantage, as nearly all of these drivers had to complete the national UK hazard perception test in order to get their driving license (it was introduced in 2002). In contrast, this may have been the first time that many of the ERDs had encountered a hazard perception test.

A second limitation lies in the low measures of internal reliability. The Cronbach's Alpha for the hazards was 0.5, below the typically acceptable level of 0.7. While low levels of internal reliability are

frequent in this field (e.g. Horswill and McKenna, 2004), it should be an aim to raise at least the reliability of the hazards, as this may improve their ability to separate the groups.

To address the potentially confounding issue of participant experience, and as an opportunity to iterate the design to improve the effects found thus far, a second study was undertaken with an iterated version of the holistic hazard test, and with more closely matched groups of drivers.

Study 2

Study 2 was designed to iterate the protocol for creating a holistic hazard test that was begun in study 1. New materials were recorded from ELVs and appropriate footage was selected and edited with the assistance of instructors and focus groups of experts. Several problems with study 1 were also addressed in the current study:

First, while it was interesting to identify differences between ERDs and control drivers in the first study, it would be more useful for stakeholders if the test was able to differentiate between different sub-groups of ERDs. Therefore, the current study compared test performance across both highly experienced, and novice emergency drivers, in addition to control drivers.

A second problem with study 1 was the discrepancy in age and general driving experience between the control and ERD groups. The current study aimed to better match control drivers with the ERDs on these key demographics.

A third problem concerns the applicability of such a test beyond the environment in which it was filmed. In study 1, the footage was captured in Nottinghamshire. As all participants were also Nottinghamshire drivers, it is possible that geographic familiarity with the roadways impacted on task performance. For instance, even control drivers may have been aware than a certain junction is likely to give rise to specific hazards, based on their personal experience of that roadway.

Beyond the possible confounding factors of route familiarity on hazard-test performance, there is also the practical issue of whether footage recorded in a specific geographic location is still relevant to emergency drivers based in a different area of the country. If a hazard test is considered to assess and develop underlying skills that are independent of route familiarity, then such tests should be relevant to all drivers regardless of context. However, if training benefit is also gained through seeing geographically specific hazards occur on the very roads you are likely to drive upon, then locally-recorded hazard tests may offer a benefit.

While the effects of route familiarity have previously been considered in relation to attention and distractibility (e.g., Young et al., 2018, Burdett, Charlton & Starkey, 2018), there are no reports of

location familiarity impacting upon hazard test performance. To investigate this, tests were created using footage from four different counties in the East Midlands of the UK (Nottinghamshire, Derbyshire, Leicestershire and Lincolnshire). Participants were recruited from all four counties and were required to undertake both a 'familiar' and an 'unfamiliar' test, allowing the familiarity hypothesis to be assessed.

Method

Participants

Eighty-three emergency response drivers were recruited from Fire and Rescue Services across the four counties of Nottinghamshire, Derbyshire, Leicestershire and Lincolnshire (Mean age = 42.35, SD = 9.81; Mean driver experience = 23.34, SD = 9.70). Thirty-nine control participants were also recruited from the four counties. Efforts were made to age-match and experience-match with the ERDs where possible. The success of this attempt can be seen in the detailed breakdown of the participants' demographics by group in Table 1.

Table 1. A detailed breakdown of the participants' demographics by county and experience group. N = number of participants (with number of females in parentheses), A = mean age of the group (with standard deviations), and E = the mean number of years that have passed since the participants gained their driving licences (with standard deviations). The greyed rows give demographics for ERDs following a reclassification according to crash risk.

	Nottinghamshire	Derbyshire	Leicestershire	Lincolnshire	Totals
Experienced	N =10 (all male)	N = 9 (all male)	N = 13 (all male)	N = 13 (3 female)	N = 45
ERDs	Age = 46.5 (6.9)	Age = 46.0 (10.1)	Age = 48.4 (4.9)	Age =41.1 (8.1)	Age = 45.4 (7.4)
	Exp. = 27.7 (6.6)	Exp. = 26.1 (8.1)	Exp. = 30.9 (5.3)	Exp. = 22.8 (7.8)	Exp. = 19.3 (10.5)
Inexperienced	N =10 (all male)	N = 8 (all male)	N = 10 (all male)	N = 10 (1 female)	N = 38
ERDs	Age = 42.6 (10.1)	Age = 38.7 (13.9)	Age = 42.8 (10.3)	Age = 31.4 (7.5)	Age = 38.9 (11.1)
	Exp. = 21.8 (8.63)	Exp. = 19.5 (14.0)	Exp. = 24.3 (10.2)	Exp. = 11.6 (4.8)	Exp. = 19.3
Control	N = 10 (2 female)	N = 10 (1 female)	N = 11 (all male)	N = 8 (1 female)	N = 39
Drivers	Age = 41.9 (14.0)	Age = 43.5 (10.1)	Age = 45.84 (13.07)	Age = 37.0 (16.4)	Age = 42.4 (13.1)
	Exp. = 22.3 (15.0)	Exp. = 25.6 (11.0)	Exp. = 24.27 (13.8)	Exp. = 16.0 (13.0)	Exp. = 22.4 (13.3)

High-Risk	*N = 9 (all male)* *Age = 45.2 (8.1)* *Exp. = 25.8 (7.6)*	*N = 11 (all male)* *Age = 41.7 (10.5)* *Exp. = 21.8 (9.9)*	*N = 11 (all male)* *Age = 44.8 (8.8)* *Exp. = 26.6 (9.0)*	*N = 12 (2 female)* *Age = 35.7 (8.4)* *Exp. = 17.4 (8.1)*	*N = 43* *Age = 41.5 (9.5)* *Exp. = 22.6 (9.3)*
Low-Risk	*N = 10 (all male)* *Age = 45.1 (8.1)* *Exp. = 23.5 (8.7)*	*N = 6 (all male)* *Age = 44.1 (10.5)* *Exp. = 25.2 (14.4)*	*N = 12 (all male)* *Age = 47.1 (7.4)* *Exp. = 29.3 (7.8)*	*N = 11 (2 female)* *Age = 38.2 (10.1)* *Exp. = 18.5 (9.3)*	*N = 39* *Age = 43.3 (10.2)* *Exp. = 24.1 (10.2)*

Design

The study employed a 2 x 3 mixed design with driver group as a between-groups factor (experienced ERDs, inexperienced ERDs, and control drivers), and route familiarity as the within-groups factor (familiar or unfamiliar). Drivers completed a test from their own county (familiar) and one from a different county (unfamiliar). Nottinghamshire and Derbyshire drivers each completed both the Nottinghamshire and Derbyshire tests, and drivers from Leicestershire and Lincolnshire each completed both the Leicestershire and Lincolnshire tests. The dependant variables were hazard perception score and accuracy of answers given to multiple-choice questions that would appear during pauses in the playback of the video clips. The order of the two tests was counterbalanced across participants.

The protocol for developing the tests

Filming

The holistic tests required video footage to be recorded from multiple cameras (Go Pro Hero 4, Silver Edition) attached to an ELV during a blue-light training run. Cameras were attached to an emergency light vehicle in a similar manner to study 1, though the camera angle of the forward-facing camera was changed to raise the horizon compared to the footage captured in study 1 (compare Figures 1 and 4).

Panel A: Normal driving

Panel B: Precursor

Panel C: Hazard

Figure 4. A typical hazardous situation from A) typical driving to B) evidence of a potential hazard (precursor) to C) clear evidence of hazard where the driver would need to slow down in order to avoid the car on the right pulling out. The hazard onset would be defined, in this instance, when the vehicle begins to pull out the junction.

Footage was collected over a 4-month period from October 2016 – January 2017 across all four counties (Nottinghamshire, Derbyshire, Leicestershire and Lincolnshire). The filming took place during several driver training, or driver refresher, courses to avoid the necessity of undertaking additional non-emergency blue-light runs beyond those required for training purposes. In total approximately 20 hours of footage was obtained. All hazardous events filmed from the ELVs arose from opportunistic on-road filming whilst the vehicles were progressing under blue lights with periodic sirens.

Early selection and initial editing of footage

One clip was selected from the footage from each county. Each clip lasted between 12 and 16 minutes long. These were initially selected based on the number of explicit hazards they contained as determined by a group of traffic psychologists. To improve upon the hazards used in study 1, we only identified hazards where the danger posed was enough to have resulted in a collision if at least one of the road users involved did not act. Following selection, the footage from the four cameras was synchronised and embedded into the graphic overlay of the car interior (as used in study 1).

Feedback from study 1 suggested that our ERDs did not like the silence of the test; even though hazard perception tests are typically silent, our drivers felt that this did not represent the distracting cacophony that accompanies emergency response driving. Accordingly, we used the audio track from the forward-facing (internal) camera to provide contextual engine and siren noise. Where the driver or instructor spoke during the film drive, the wave patterns associated with the speech were isolated from engine/siren noise and deleted in audio editing software.

Focus groups

Three focus groups were held with seven FRS driving instructors from the four participating counties. Our initial selections of footage were played to these groups and the instructors stopped playback at points where they believed a hazard had occurred, or where they thought there was a good opportunity to ask an MCQ. Through such discussion, the group helped refine the list of early-selection *a priori* hazards, rejecting some and including ones that had been missed or disregarded by the traffic psychologists. The instructors also identified many critical decision points and crucial observations that could act as multiple-choice questions and assisted with the wording of options for the MCQs.

The final hazards

The final number of hazards in each clip varied due to the nature of the county and the type of driving that we were able to capture on camera. Scoring windows were defined from the start of hazard onset (typically when another road user begins a trajectory that would end in collision with the film car

unless adjustment is made) to hazard offset (when the film car is past the danger). For instance, one hazard contained a young pedestrian who stepped out from between parked cars as the film car approach. He was looking the wrong way as he stepped out (and presumably had headphones underneath his hood, which prevented him from hearing the sirens). As the film car approached, the pedestrian turned his head, caught sight of the ELV, and jumped back onto the pavement. Hazard onset was identified as when the head of the pedestrian was first visible over the roofs of the parked cars, and offset was triggered when the pedestrian jumped backwards. Any button press within this temporal window was considered a hit. Response times were converted in scores from 0-5 following the scoring method used in study 1.

The final multiple-choice questions

Four different types of questions were edited into the clips. The nature of the footage across the four different tests dictated the actual questions, resulting in the absolute numbers question types varying across the clips.

- What happens next? (WHN): the clip stops and is occluded by a highly blurred final frame just as a hazard onsets and drivers are asked, "What happens next?" These questions addressed different hazards to those that required a speeded-button response. No information was available in the blurred frame to allow identification of the answer. Participants then choose the correct answer out of 4 options before the clip continued;
- What is the main source of hazard? (MSoH): While WHN questions had to have an actual event that subsequently materialised, there were many other situations where something *could have* happened but did not. To avoid wasting some of these excellent opportunities for assessing and training advanced situation awareness, we asked drivers to choose the main source of a potential hazard following occlusion of the clip. This was akin to asking, "What is most likely to happen next?" Following the selection of one of the four options, the clip would continue. Drivers did not get any feedback as to whether their answer was correct, as no hazard actually occurred. Correct answers were provided by the instructors during the focus groups. An MSoH example is given in Figure 5.
- Observation questions (OBS): The focus groups were keen to include general observation questions to ensure that participants were attending to relevant aspects of the scene (e.g. "What is the posted speed limit in this area"). Questions again followed an occlusion; with the clip resuming once one of four options was selected.
- What would you do next? (WWYDN): Initially these questions were to follow an occlusion as with all other questions, however the instructors advised that this would be too difficult.

Accordingly, the video was paused and left visible for 3 seconds with the question at the top of the screen. The image was then replaced with the four options. These questions typically involved decisions to contravene traffic rules. Some of the options advocated ostensibly more risky manoeuvres, while others were more risk-averse. The correct answers to these questions were split across cautious and progressive answers as determined by the instructors in the focus groups. Following selection of an option, the clip continued, and participants could see whether their choice was the same as that of the film car driver (though they were told that the actions of the film driver might not have been correct in some instances).

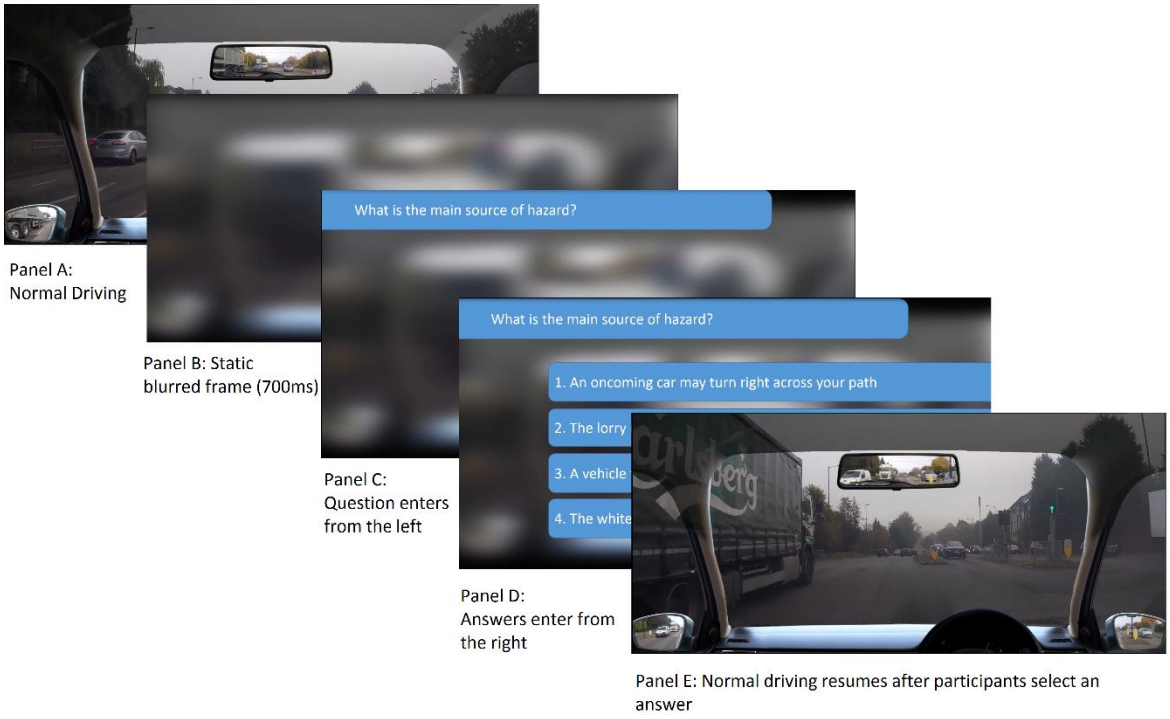

Figure 5. Screen shots from a clip with an MSoH question. At certain points during normal driving (A) an event would occur, and the video would pause and blur preventing access to the majority of the visual information in the scene (B). A question would then appear at the top (C), followed by 4 answers (D). Normal driving would resume after an answer was selected using the appropriate number key on a keyboard (E).

The Completed Tests

The final tests contained an average of 11 hazards requiring a button press, 5 WHN questions, 4 MSoH questions, 3 OBS questions and 6 WWYDN questions. They were embedded into Eprime 3 control software in order to present the clips to participants and collect responses.

Additional stimuli and apparatus

Lenovo Thinkpad Yoga 370 laptops were used to administer the tests. Each had a screen size of 13.3" and ran the tests at a resolution of 1920 x 1080. A mouse was used for participants to click when they saw hazards and the buttons 1,2,3 and 4 on the keyboard was used to select answers for the multiple-choice questions. Headphones were provided to ensure focus on the task, and to avoid disruption to other participants in multiple-participant testing sessions.

A demographics questionnaire was designed to capture basic details including age, driving experience (of both normal driving and emergency response driving, where appropriate) and the number of collisions that drivers could remember over the course of their driving history.

In addition, a variant of the Driver Behaviour Questionnaire (Parker et al., 1995) was given to our participants, but as no group effects were found this will not be referenced further.

Procedure

Fire personnel were tested in their respective fire stations while on shift. Control participants were tested either in the laboratory, or field locations across the four counties (Lincoln library, Derby Costa, Leicester University Student Union, etc.). Each participant was asked to fill in a demographics questionnaire and then complete two tests, before finishing with the Driver Behaviour Questionnaire. The procedure was otherwise identical to study 1.

Results

Data cleaning

A total of 122 participants were tested across all groups. Six participants were removed from all analyses (1 Derby experienced, 1 Derby inexperienced, 1 Leicester control, 1 Leicester inexperienced and 2 Nottingham controls), as they did not respond to any of the hazards. As such, it was deemed that they were not sufficiently engaged with the task or had misunderstood the instructions. A further 3 participants were removed due to data loss caused by equipment failure (1 Derby inexperienced, 1 Nottingham inexperienced and 1 Lincoln experienced). Finally, one of the hazards selected in the Nottinghamshire clip failed to receive any responses within its scoring window. This hazard was removed from all analyses. One other hazard (from Lincolnshire) only had 8% of drivers make a correct hazard response. As this was below three standard deviations from the mean percentage of successful respondents across all hazards, this event was also removed from all subsequent analyses.

Hazard scores across driver experience groups

As in study 1 we followed the method used by the DVSA to score the national hazard perception test, the scoring windows for each hazard were split into 5 even sections. As different tests contained different numbers of hazards, we compared the mean score per hazard (from zero to 5) across all groups. The average score per hazard across driver groups and the familiarity factor can be viewed in Table 2.

Table 2. The average hazard scores and multiple-choice scores across driver groups and the familiarity factor (with standard deviations). The greyed-out rows provide means associated with our ERDs following reclassification as either high-risk or low-risk.

	Familiar		*Unfamiliar*	
	Hazard Score	MCQ Score	Hazard Score	MCQ Score
Experienced	2.54 (1.11)	54.62% (10.56)	2.60 (1.06)	52.15% (13.35)
Inexperienced	2.46 (1.04)	54.62% (10.56)	2.73 (0.93)	51.45% (13.79)
Control	2.11 (0.98)	42.29% (14.28)	2.22 (1.15)	42.31% (11.39)
High-Risk	*2.70 (0.98)*	*56.92% (11.05)*	*2.55 (1.03)*	*53.77% (13.67)*
Low-Risk	*2.27 (1.13)*	*54.08% (13.02)*	*2.73 (0.94)*	*49.86% (13.14)*

A 2 x 3 Analysis of Variance (ANOVA) compared driver groups across familiar and unfamiliar tests. Though the main effect of driver group only approached significance ($F(2, 110) = 2.74$, MSE = 2.19, $p = 0.07$), the planned Helmert contrasts revealed that all fire car drivers scored significantly more points than control drivers (M = 2.58 vs M = 2.16 $p = 0.02$), though there was no different between the novice and experienced ERD groups. The main effect of familiarity did not reach significance, $F(1, 110) = 1.95$, MSE = 1.21, $p = 0.17$. There was also no suggestion of an interaction between the two factors, $F(2, 110) = 0.38$, $MSE = 0.24$, $p = 0.68$. Cronbach's alpha for the response rates to the hazards in each county are presented in Table 3.

Table 3. Cronbach's alpha for responses to hazards for each test in study 2, with Study 1 included for comparison.

	Study 1		Study 2				
			Nottingham-shire	Derbyshire	Leicestershire	Lincolnshire	Total
Cronbach's Alpha	0.5		0.74	0.35	0.83	0.80	0.68

In each test, participants were presented with multiple-choice questions (MCQs) during the video clip (17 MCQs for Derby test, 23 MCQs for Nottingham test, 16 MCQs for Leicestershire test and 18 MCQs for the Lincoln test). KR20 statistics for the MCQs were once again low, with overall ratings ranging from 0.25 to 0.51; better than found in study 1, but still not close to an acceptable level of 0.7. Overall percentage accuracies for all participants for each of the tests they completed (familiar and unfamiliar) can be viewed in Table 2. The resultant data were then subjected to a 2 x 3 ANOVA, which yielded a significant main effect of driver group, $F(2, 110) = 20.20$, $MSe = 3183.55$ $p < 0.001$. Helmert contrasts revealed that control drivers ($M = 42.30\%$) were significantly less accurate in answering than the questions than all ERDs (both inexperienced and experienced, 54% and 53%, respectively, $p < 0.001$); see Figure 6.

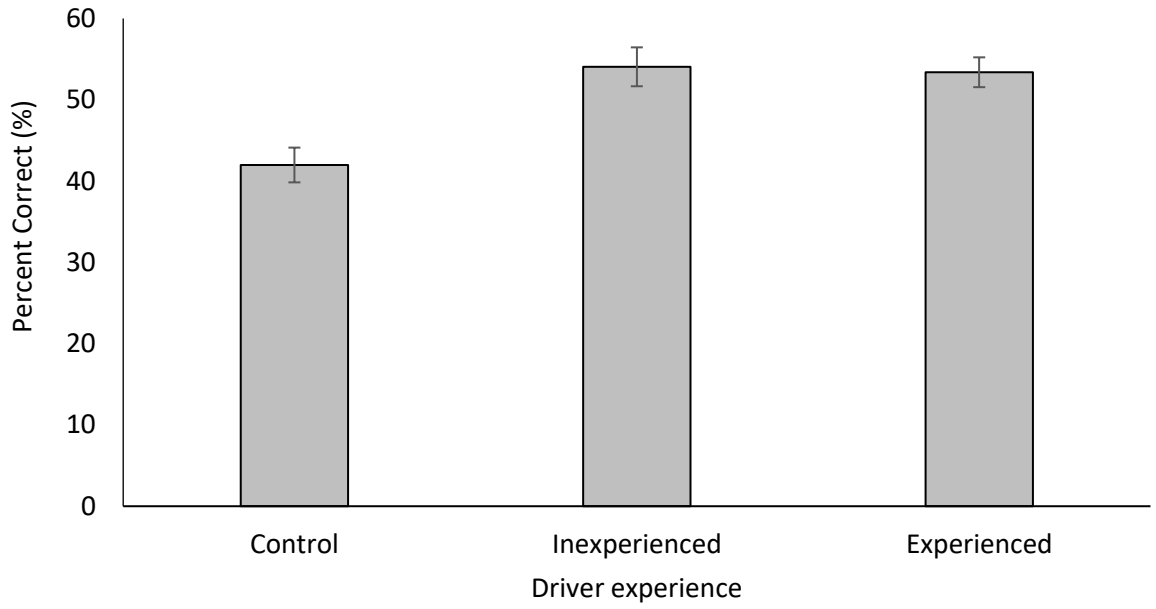

Figure 6. The mean percentage of correct answers to the multiple-choice answers given by the three driver groups (with standard error bars).

To assess the contribution of individual questions to this effect, the percentage of experienced and inexperienced ERDs and control drivers who responded correctly to each question for each county was also charted (see Figure 7). As in study 1, this graph indicates that not all items contributed towards the overall superiority for ERDs on the MCQs. While many items suggest greater accuracy for ERDs over control participants, there are some items where the control participants score higher than the ERDs. Despite this, in the majority of items in all counties, the ERDs outperform the control drivers.

Re-analysing hazard scores according to driver risk

Whilst the current results are in line with the findings from the study 1, which found advantages for ERDs over control participants, there were no experiential differences between the two ERD groups. We found similar results in our previous study of fire-appliance drivers' response times to hazards (Crundall & Kroll, 2018): while all FA drivers out-performed control drivers, there was no difference between experienced and inexperienced FA drivers. Crundall and Kroll (2018) however found low-risk fire-appliance (FA) drivers to outperform high-risk FA drivers on a hazard prediction test when split according to high and low risk (based on previous collision ratings). Might this pattern also occur with our emergency drivers? To determine whether the current test is perhaps more sensitive to driver risk rather than experience, we recoded all ERDs according to risk. All participants with 1 or fewer collisions in the entirety of their driving life span were coded as *low-risk* and those who had had 2 or more collisions were coded as *high-risk* (following Das et al., 2015). One participant was removed, as they did not provide collision data, so we could not code them as either high or low-risk.

The average scores per hazard across driver groups and the familiarity factor can be viewed in Table 2. A 2 x 3 ANOVA (driver group x familiarity) was conducted on these scores. Once again, the main effect of driver group approached significance, $F(2, 109) = 2.72$, $MSE = 2.14$, $p = 0.07$, while planned Helmert contrasts revealed that all ERDs were significantly faster than control drivers (M = 2.56 vs M = 2.16, $p = 0.03$). There were no differences between low-risk and high-risk ERDs ($p = 0.53$), and the main effect of familiarity was not significant, $F(2, 109) = 1.81$, $MSE = 1.08$, $p = 0.181$. The interaction between the factors approached significance, $F(2, 109) = 2.96$, $MSE = 1.76$, $p = 0.056$. Post-hoc corrected pairwise comparisons revealed that low-risk drivers were faster to detect hazards on unfamiliar roads than familiar roads ($M = 2.73$ vs. $M = 2.27$, $p < 0.001$). Viewing the means in Figure 8, suggests that the crucial effect driving the interaction is an apparent drop in performance for low-risk drivers when viewing the test containing familiar locations.

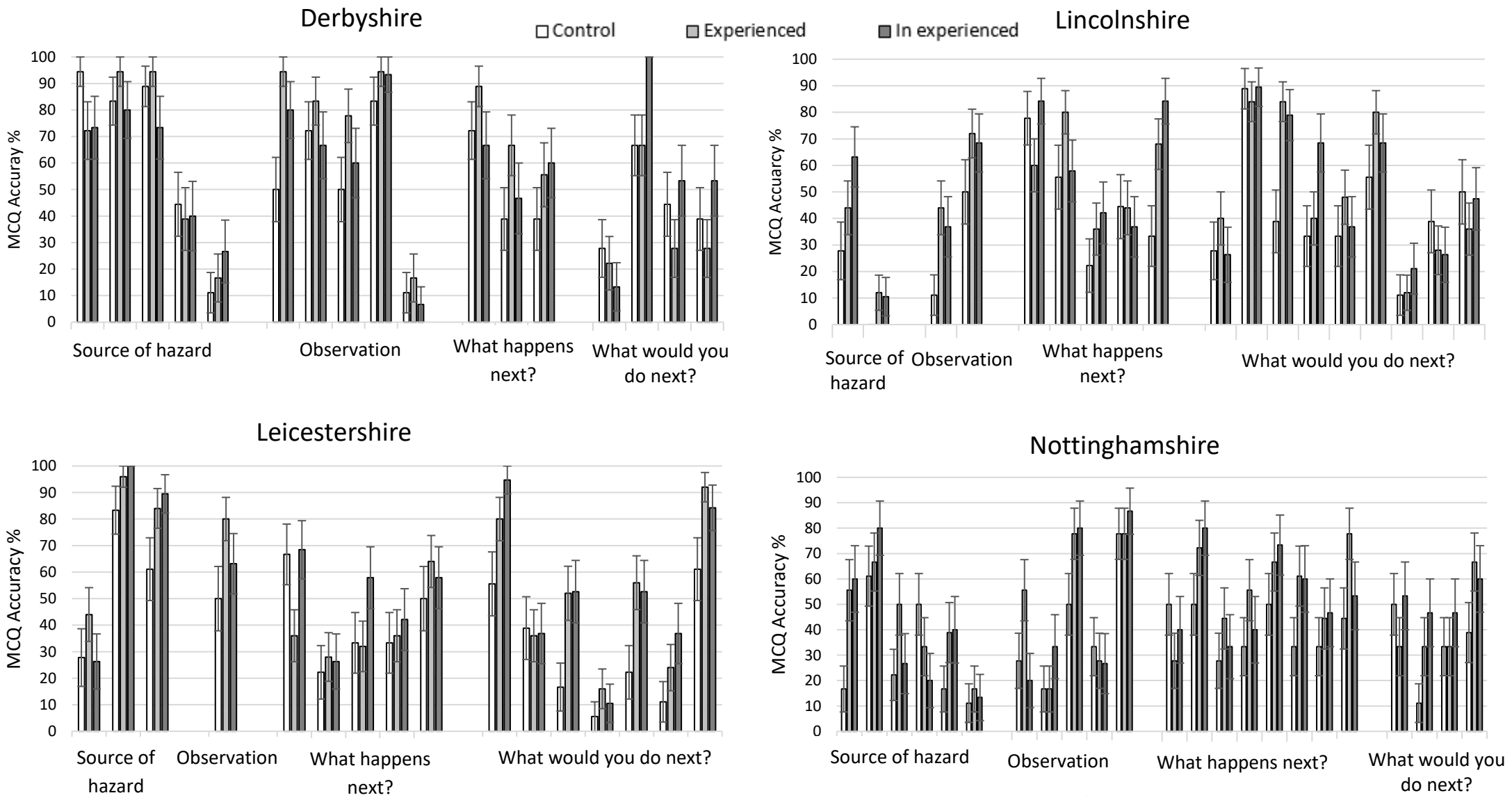

Figure 7. The percentage of participants who responded correctly to each of the multiple-choice questions (MCQs) in each county. Standard error bars are added.

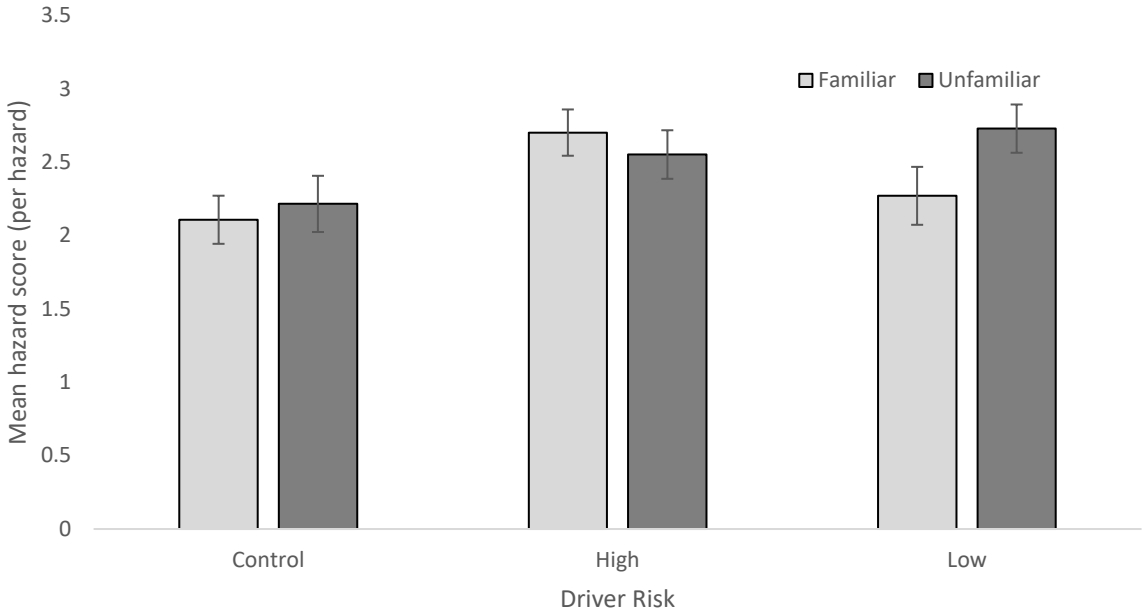

Figure 8. The mean score per hazard across and familiarity and driver groups, with ERDs re-classified according to crash risk (with standard error bars). The minimum score possible was zero and the maximum was 5.

Re-analysing MCQ scores according to driver risk

Percentage accuracies for all participants were calculated for each of the tests they completed (familiar and unfamiliar; see Table 2). A 2 x 3 ANOVA (driver group x familiarity) revealed a significant main effect of driver group, $F(2, 110) = 21.99$, $MSE = 3406.03$, $p < 0.001$. The pattern of means was almost identical to that found when ERDs were classified according to experience (shown in Figure 6), with both ERD groups out-performing the control drivers, though they did not differ from each other (as confirmed by planned contrasts).

Discussion

The pattern of results from study 2 follows a similar pattern to that of study 1. There is evidence that ERDs are faster to respond to hazards than control drivers (at least at the level of planned contrasts) and score more highly on the multiple-choice questions. This confirms the sparse literature that suggests ERDs are more effective at hazard perception than control drivers (Johnston & Scialfa, 2016; McKenna and Crick, 1991). This study is one of only three that have demonstrated this with task-specific footage (see Crundall & Kroll, 2018; Crundall et al., 2003, 2005). Furthermore, the measures of internal consistency (both KR20 and Cronbach's Alpha) were improved over study 1, with three of our four tests producing alphas of over 0.7 for the hazards.

The two experiential groups of ERDs did not differ in terms of their hazard perception or MCQ responses. This was also observed in the fire-appliance (FA) hazard perception test (Crundall and Kroll, 2018). While FA drivers' response times to hazards were found to differ to those of control participants, there was no difference between sub-groups of FA drivers based on their years of experience driving a fire appliance (although eye-movement measures were able to discriminate between the two FA driver groups).

Accordingly, we re-categorized our participants according to frequency of recalled collisions. Crundall and Kroll (2018) found FA driver risk to be a more effective discriminator of FA groups, and it was rationalised that this may also be the case within the current data. Splitting our emergency drivers into high and low risk groups (where *high* was defined as having two or more collisions in recent memory) also failed to reveal a main effect difference in hazard scores. However, an interaction between ERDS group (high risk vs. low risk) and route familiarity was observed, albeit with the results painting an odd picture: those ERDs who reported fewer collisions appeared to perform poorly on a local test based on familiar roads.

Why might this negative effect of familiarity have occurred? There is published evidence that shows drivers to pay less attention to safety-critical aspects of the scene on familiar roads (Young et al., 2017) and that this may partially underlie the 'close-to-home' effect in crash statistics (Burdett et al., 2017). We reject this as a possible explanation of the current results for the following reasons. First, there is no obvious reason why this interpretation of the familiarity effect would impact on our low-risk drivers, but not the high-risk drivers. Secondly, when professional drivers are assessed by traffic psychologists, they tend to try their hardest, at least over short periods of time. The deleterious effect of route familiarity reported by Young et al. (2017) was found after a driving instructor had driven the same route on 28 occasions over a series of weeks, after which time, even the most dedicated road safety professional is likely to find their concern over assessment waning. In contrast, the current experiment rarely took more than 45 minutes to complete in a single session (and the clips took less than 30 minutes of this time). We strongly suspect that our ERDs would have maintained a high-level of motivation throughout this short testing period.

An alternative explanation comes from the possible negative effects of 'surprise recognition'. We found that several ERDs would spontaneously exclaim, "I know this road!" Unbidden thoughts cued by familiar locations can overload working memory and degrade current processing of visual stimuli (i.e. drivers may temporarily lose concentration). In addition, recognition of locations may have influenced their visual scanning. For instance, emergency drivers, familiar with the route, may inspect

parts of the world that are particularly memory-evoking (e.g. a statue) or look at things that are incongruent with their memory (e.g. a new shop frontage).

Again, however, we must question why this effect might be greater for low-risk ERDSs? One possibility is that low-risk drivers rarely get the chance to look at scenery when driving (because they may be more dedicated to searching for hazards). Given the opportunity to view familiar roads in a completely safe environment (after all, they are not really driving during this test), their curiosity may intermittently override the primary task.

One final influence of location recognition is that drivers may use their knowledge of the area to help predict hazards. This may have advantages in very specific situations (e.g. knowledge of a hidden road entrance may help one spot a car as it emerges). However, given the degrees of freedom in naturally occurring hazards, it is more likely that while one is focusing on the hidden road entrance, a completely different hazard may be missed. Highly motivated, low-risk emergency drivers may essentially *try too hard,* erroneously using prior knowledge to restrict their search for hazards.

While the data cannot discriminate between these post-hoc rationalisations, the results suggest that a non-familiar location might elicit the best performance from all participants. Certainly, unfamiliar locations do not unfairly hinder ERDs performance. On this basis, it can be suggested that hazard tests do not need to be filmed in the locations in which they are to be used. A Nottinghamshire-based test should be perfectly adequate for Leicestershire drivers, for instance. Generalisability of clips across different geographic training locations removes the need for bespoke tests to be created for individual locales, increasing the economic viability of introducing hazard perception tests as a national training and assessment tool for emergency service drivers.

It should be noted, however, that there are extreme locations that will still benefit from some level of bespoke tailoring. Our clips span a variety of road types from urban and suburban, to arterial and rural. While this mix may fit most locations in the UK, some extremely urban and extremely rural locations may benefit from location-specific hazard tests. The highlands of Scotland provide drivers with a very different mix of hazards to those faced by drivers in the centre of London, and neither may be best served by a selection of tests filmed across varying roadways in the East Midlands.

The lack of differences between the experienced and novice ERD groups should also be addressed. It is not uncommon for hazard perception studies to find no differentiation between driver groups of varying experience (e.g., Chapman & Underwood, 1998; Crundall et al., 1999, 2002; Groeger et al., 1998), and this has been attributed to several underlying reasons, such as criterion bias (Crundall et al., 2003, 2005). Equally however, it is possible that our novice ERDs are relatively good at the task.

They are all recently-trained response drivers as well as highly-experienced car drivers and will have been exposed to many emergency response situations as a passenger in an ELV or a fire appliance prior to training.

General Discussion

The aim of this research was assess the potential of a holistic hazard test for the emergency services, which combined traditional hazard perception responses (press a button as quickly as possible to a hazard with multiple-choice questions (MCQs) designed to assess situation awareness and decision-making. This was done using naturalistic footage captured from an emergency light vehicle (ELV) across several counties in the East Midlands on blue-light training runs (a method only used twice previously by Crundall & Kroll, 2018 and Crundall et al., 2003, 2005).

Across two studies, 5 tests and 162 participants, the results suggest that ERDs have a tendency to spot and respond to hazards faster than controls, and that they score more highly on multiple-choice questions designed to probe hazard awareness and observation skills. This provides a potentially useful starting point for developing assessment and training tools for emergency service drivers that can be used to supplement on-road training.

We did not find, however, any indication that our tests can find substantial group differences between different sub-groups of ERDs. The only difference noted between our sub-groups was that low-risk drivers appeared to perform poorly on the familiar route. We suggested that low-risk drivers may try to use local knowledge to gain an advantage in predicting the dangers, though in several cases the actual hazards might be better responded to by reading the road than applying geographically specific prior knowledge.

On this basis we suggest that our tests might not be suitable for differentiating drivers who have been trained and already have experience of response driving (whether novice or experienced). However, the consistent differences between ERDs and control drivers suggest that this could be a useful tool at the point of first assessment for inclusion on an ERDS course, or within the first few training sessions. We also recommend that drivers are assessed on non-local tests, or are at least warned against the problems of potential memory-based distraction. This has the benefit however that one test can have much wider applicability, as lack of familiarity with the roadways does not appear to have a deleterious effect on performance. This makes the development of ERD hazard tests economically viable, as ‘one size (or location) fits all’. As noted above, this is not to say that certain. Indeed, a familiar route could interfere with the test’s ability to assess drivers’ hazard skills.

The method has combined a number of innovations including using emergency response footage, the inclusion of MCQs items embedded into the video playback, the use of a single clip rather than a selection of short clips, and even the inclusion of sound in study 2. The study did not set out to test the impact of each of these innovations in isolation, so we cannot conclude at this stage whether the individual innovations were successful, though it will be possible for future studies to unpack these separate components. For instance, it would be interesting to identify whether the inclusion of MCQs affect hazard perception response times. The MCQs added an unpredictable secondary task, which presumably placed additional demands on attentional resources. It can be argued that this is closer to the situation drivers face on the road, where they may be unable to devote all their attention to spotting the next hazard.

We must however acknowledge the limitations of the current study. First, there is more development required. Hazards and probe questions can be improved in order to separate groups on the basis of performance and to improve internal reliability. Once content has been improved, more detailed analyses need to be undertaken to assess how successful these tests might be at categorising drivers as skilled or less-skilled at emergency response driving at an individual level. We also need to explore the most appropriate methods for developing training materials from such tests. At the very least, drivers will be keen to see whether their answers were correct. While some of our probe questions (e.g. What Happens Next?) inevitably provide feedback when the clip resumes playback, other questions did not. Future research must assess the impact of feedback and other training options on hazard performance in these tests, and also on real-world response driving.

Nonetheless, the results are a first step in moving towards off-road training and assessment supplements for emergency response drivers that are likely to be cost-efficient. Given the cost and risk of taking trainees out on real blue-light training runs, there is potential benefit to be gained by providing trainees with the opportunity to view typical hazards prior to switching on their sirens for the first time.

Acknowledgements

This research was funded by the Fire Service Research and Training Trust (FSRTT). We would like to thank Craig Parkin, Viv Brunsden, Mark Jarman, Paul Foster, Ivan Hayes, Trevor Neale, Martin Lemmon and all the fire service personnel who assisted in the development of the tests.

References

Becker, L. R., Zaloshnja, E., Levick, N., Li, G., and Miller, T. R. (2003). Relative risk of injury and death in ambulances and other emergency vehicles. *Accident Analysis and Prevention, 25, 6,* 941-948.

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.

Braitman, K. A., Kirley, B. B., McCartt, A. T., & Chaudhary, N. K. (2008). Crashes of novice teenage drivers: Characteristics and contributing factors. Journal of Safety Research, 39(1), 47-54.

Burdett, B. R., Starkey, N. J., & Charlton, S. G. (2017). The close to home effect in road crashes. Safety science, 98, 1-8.

Burdett, B. R., Charlton, S. G., & Starkey, N. J. (2018). Inside the commuting driver's wandering mind. Transportation research part F: traffic psychology and behaviour, 57, 59-74.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81.

Castro, C., Padilla, J. L., Roca, J., Benítez, I., García-Fernández, P., Estévez, B., Lopez-Ramon, M., F., Crundall, D. (2014). 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., & 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. E., & Underwood, G. (1998). Effects of experience and processing demands on visual information acquisition in drivers. Ergonomics, 41(4), 448-458.

Crundall, D., Andrews, B., Van Loon, E., & Chapman, P. (2010). Commentary training improves responsiveness to hazards in a driving simulator. Accident Analysis & Prevention, 42(6), 2117-2124.

Crundall, D., Chapman, P., 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.

Das, S., Sun, X., Wang, F. and Leboeuf, C., (2015). Estimating likelihood of future crashes for crash-prone drivers. Journal of traffic and Transportation engineering, 2, 3, 145-157.

Deery, H. A. (2000). Hazard and risk perception among young novice drivers. Journal of Safety Research, 30(4), 225-236.

Drummond, A. E. (2000, June). Paradigm lost! Paradise gained? An Australian's perspective on novice driver safety. In Novice Drivers Conference, 2000, Bristol, United Kingdom.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, , 32(2) 97-101.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(1), 32-64.

Fisher, D. L., Pollatsek, A. P., & Pradhan, A. (2006). Can novice drivers be trained to scan for information that will reduce their likelihood of a crash? Injury Prevention : Journal of the International Society for Child and Adolescent Injury Prevention. 12(Suppl 1), i25–i29.

Garay, L., Fisher, D. L., & Hancock, K. L. (2004). Effects of driving experience and lighting condition on driving performance. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48(19) 2290-2294.

Garay-Vega, L., & Fisher, D. L. (2005). Can novice drivers recognize foreshadowing risks as easily as experienced drivers. Proceedings of 3rd International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design. Rockport, Maine: USA,

Groeger, J. A., & Chapman, P. (1996). Judgement of traffic scenes: The role of danger and difficulty. Applied Cognitive Psychology, 10(4), 349-364.

Hancock, P. A., & De Ridder, S. N. (2003). Behavioural accident avoidance science: understanding response in collision incipient conditions. Ergonomics, 46(12), 1111-1135.

Horswill, M. S., & McKenna, F. P. (2004). Drivers' hazard perception ability: Situation awareness on the road.

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.

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. Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 63, P212–P218.

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

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 & Prevention, 95, 91-96.

Lajunen, T., & Summala, H. (2003). Can we trust self-reports of driving? Effects of impression management on driver behaviour questionnaire responses. Transportation research part F: traffic psychology and behaviour, 6(2), 97-107.

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

Matthews, M. L., & Moran, A. R. (1986). Age differences in male drivers' perception of accident risk: The role of perceived driving ability. Accident Analysis & Prevention, 18(4), 299-313.

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

Malaterre, G. (1988).How to estimate feasible maneuvers when driving a car in an emergency situation. Traveil Humain, 51(3), 211-227.

Maycock, G., Lockwood, C., & Lester, J. F. (1991). The accident liability of car drivers.

Mayhew, D. R. (2003). The learner's permit. Journal of Safety Research, 34(1), 35-43.

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

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.

McKenna, F. P., & Horswill, M. (1999). Hazard perception and its relevance for driver licensing. IATSS Research, 23(HS-042 879)

Mueller, A. S., & Trick, L. M. (2012). Driving in fog: The effects of driving experience and visibility on speed compensation and hazard avoidance. Accident Analysis & Prevention, 48, 472-479.

Parker, D., Reason, J. T., Manstead, A. S., & Stradling, S. G. (1995). Driving errors, driving violations and accident involvement. Ergonomics, 38(5), 1036-1048.

Pelz, D. C., & Krupat, E. (1974). Caution profile and driving record of undergraduate males. Accident Analysis & Prevention, 6(1), 45-58.

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.

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

Pradhan, A. K., Hammel, K. R., DeRamus, R., Pollatsek, A., Noyce, D. A., & Fisher, D. L. (2005). Using eye movements to evaluate effects of driver age on risk perception in a driving simulator. Human factors, 47(4), 840-852.

Quimby, A., Maycock, G., Carter, I., Dixon, R., & Wall, J. (1986). Perceptual Abilities of Accident Involved Drivers. TRRL Laboratory Report 1004. Crowthorne: Transport and Road Research Laboratory.

Quimby, A., & Watts, G. (1981). Human Factors and Driving Performance. TRRL Laboratory Report 1004. Crowthorne: Transport and Road Research Laboratory.

Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: a real distinction?. Ergonomics, 33(10-11), 1315-1332.

Renge, K. (1998). Drivers'hazard and risk perception, confidence in safe driving, and choice of speed. IATSS Research, 22(2), 103-110.

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

Scialfa, C. T., Deschênes, M. C., Ference, J. D., Boone, J., Horswill, M. S., & Wetton, M. (2012). Hazard perception in older drivers. International journal of human factors and ergonomics, 1(3), 221-233.

Spicer, R.A. (1964). Human factors in traffic accidents. Final Report (Research Grant No AC-55). United States Public Health Service.

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

Underwood, G., Crundall, D., & Chapman, P. (2002). Selective searching while driving: The role of experience in hazard detection and general surveillance. Ergonomics, 45(1), 1-12.

Underwood, G., Ngai, A., & Underwood, J. (2013). Driving experience and situation awareness in hazard detection. Safety Science, 56, 29-35.

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

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.

Wallis, T. S., & Horswill, M. S. (2007). Using fuzzy signal detection theory to determine why experienced and trained drivers respond faster than novices in a hazard perception test. Accident Analysis & Prevention, 39(6), 1177-1185.

Watts, G., & Quimby, A. (1979). Design and Validation of a Driving Simulator for use in Perceptual Studies,

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

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

Wetton, M., Horswill, M., 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.

Young, A. H., Mackenzie, A. K., Davies, R. L., & Crundall, D. (2018). Familiarity breeds contempt for the road ahead: The real-world effects of route repetition on visual attention in an expert driver. Transportation research part F: traffic psychology and behaviour, 57, 4-9.