Running Header: CROSS CULTURAL HAZARD PERCEPTION AND PREDICTION

A comparison of hazard perception and hazard prediction tests across China, Spain and the UK

Petya Ventsislavova, David Crundall and Thom Baguley

Nottingham Trent University

Candida Castro, Andrés Gugliotta and Pedro Garcia-Fernandez

University of Granada

Wei Zhang, Yutao Ba, and Qiucheng Li

Tsinghua University

Author Note

Petya Ventsislavova, David Crundall, Thomas Baguley, Department of Psychology, School of Social Sciences, Nottingham Trent University, UK.

Candida Castro, Andrés Gugliotta, Pedro Garcia-Fernandez, CIMCYC, Mind, Brain and Behaviour Research Centre, University of Granada, Spain.

Li Qiucheng, Yutao Ba, & Wei Zhang, Department of Industrial Engineering, Tsinghua University, Beijing, China. Yutao Ba is now based at IBM China, Zhongguancun Software Park 19, Haidian, Beijing, China, 100094.

This research was supported, in part, by funding from Santander. Correspondence concerning this article should be directed to David Crundall (david.crundall@ntu.ac.uk) or Petya Ventsislavova (petya.petrova@ntu.ac.uk)

Abstract

Hazard perception (HP) is the ability to spot on-road hazards in time to avoid a collision. This skill is traditionally measured by recording response times to hazards in video clips of driving, with safer, experienced drivers often out-performing inexperienced drivers. This study assessed whether HP test performance is culturally specific by comparing Chinese, Spanish and UK drivers who watched clips filmed in all three countries. Two test-variants were created: a traditional HP test (requiring timed hazard responses), and a hazard prediction test, where the film is occluded at hazard-onset and participants predict what happens next. More than 300 participants, across the 3 countries, were divided into experienced and inexperienced-driver groups. The traditional HP test did not discriminate between experienced and inexperienced drivers, though participant nationality influenced the results with UK drivers reporting more hazards than Chinese drivers. The hazard prediction test, however, found experienced drivers to out-perform inexperienced drivers. No differences were found for nationality, with all nationalities being equally skilled at predicting hazards. The results suggest that drivers' criterion level for responding to hazards is culturally sensitive, though their ability to predict hazards is not. We argue that the more robust, culturally-agnostic, hazard prediction test appears better suited for global export.

Keywords: hazard perception, hazard prediction, driving safety

Introduction

Countries with low levels of on-road injuries and fatalities should look to export their most successful safety initiatives to countries with higher rates of traffic collisions. The UK hazard perception test has been heralded as one of the most successful initiatives of recent times, having been associated with significant reductions in certain types of traffic collision after incorporation into the UK licensing procedure in 2002 (Wells et al., 2008). But is it suitable for export? This depends on whether there are cultural differences in the way people respond to the hazard perception test. This paper reports a study that assessed participant responses to a traditional hazard perception test across three countries, and found it wanting. A second study, however, used a variant on the traditional methodology which we found to be more suitable for testing driver skill across geographical borders.

Traffic death and injury as a global problem

Injuries and fatalities arising from traffic collisions are a global problem. The World Health Organization (2015) estimates the number of global fatalities due to traffic collisions to be 1.25 million, with up to 90% of these occurring in low to middle-income countries. Currently road traffic collisions are the 8th leading cause of death in the world, but are predicted to rise to the 5th leading cause by 2030 unless a concerted effort is made to avert this disaster. In response to this growing problem, the United Nations declared a Decade of Action for Road Safety which began in 2011, with the aim of first stabilising and then reducing road traffic fatalities and injuries by 2020. With over 100 countries pledged to assist, the Decade of Action is focused upon ‘5 pillars’ of road safety: road safety management, developing safer roads, developing safer vehicles, developing safer road users, and improving emergency responses to incidents. One key aspect of this is the technological promise of automated vehicles, but the Institute of Electrical and Electronics Engineers (http://www.ieee.org/about/news/2012/5september_2_2012.html) estimates that we will have to wait until 2040 before 75% of global driving stock is automated. Even if this ambitious target is met, millions will die before we reach this point without a wider range of safety initiatives, and it will be the low and middle-income countries that will continue to bear the brunt of this automotive pandemic.

When countries are compared on the number of road fatalities, accounting for population size, a handful of European countries typically dominate the safest spots at the top of the table (e.g. Sweden, Norway, Denmark, Switzerland, the UK, and the Netherlands; OECD/ITF, 2015; WHO, 2015). Thus it behoves researchers in these countries to identify which of their own safety initiatives contribute significantly to their national safety record, and to assess whether these interventions are suitable for export to other countries who may benefit in their own attempts to reduce on-road injuries and fatalities. In the field of traffic and transport psychology this will typically involve reviewing the impact and suitability of training, education, enforcement and assessment initiatives, in support of the UN's fourth pillar: developing safer road users.

In the UK a number of initiatives have been implemented over the decades, with many of these resulting in changes to the licensing procedure and to the rules of the road. These changes include the launch of the pass-plus scheme (a post-license training qualification launched in 1995), the introduction of a computerised touch-screen theory test to the licensing procedure (2000), a ban on the use of hand-held mobile phones while driving (2003), and the introduction of an eco-safe driving element to the driving test (2008), along with a section of independent driving (2010), where the learner must navigate by themselves for 10 minutes. New changes to the driving test are also being currently considered (including the use of satnav). One of the most influential changes to driving in the UK has been the introduction of the hazard perception (HP) test to the driving test (in 2002). This paper will focus on whether this test is suitable for export to other driving cultures.

Developed by traffic and transport psychologists, the UK HP test presents learner drivers with a series of video clips (updated to computer-generated imagery in 2015) filmed from the driver's perspective. Over 100,000 learner drivers sit this test every month in government offices throughout the country and must achieve a pass mark before they are allowed to take the on-road driving test. They are required to press a button whenever they spot a hazard that might cause the film-car to have a collision. The faster one presses a button in response to a hazard, the more points are awarded. Participants can score a maximum of 5 points per hazard, dependant on the speed of the response. Across 15 hazards, learners must achieve a pass-mark of 44 out of 75. The rationale for the introduction of this test lies upon the assumption that the faster one spots and responds to a hazard in the test is positively related to one's likelihood of avoiding a crash, and that by introducing such a test it will keep the worst drivers off the roads, while also encouraging driving instructors to focus more on the higher-order cognitive skill of hazard perception.

The development of the hazard perception test (HP test) has been supported by numerous studies that have demonstrated the ability to discriminate between collision-involved and collision-free drivers (e.g. Pelz and Krupat, 1974; Watts and Quimby, 1979; McKenna and Crick, 1991), and between novice and highly-experienced drivers (where the former are typically over-represented in collision statistics; Renge, 1998; Wallis & Horswill, 2007; Horswill et al., 2008; Deery, 1999; Pradhan, Pollatsek, Knodler & Fisher, 2009). A few studies have even reported that performance on a hazard perception test can predict whether a driver will be involved in a future collision (Boufous et al., 2011; Drummond, 2000; Horswill, Hill and Wetton, 2015). These studies support the assumption that poor levels of hazard perception skill are related to a higher likelihood of having a crash. Furthermore, there are many studies that have demonstrated that performance in detecting hazards can be trained (e.g. Castro et al., 2016; Chapman, Underwood and Roberts, 2002; Horswill, Taylor, Newnam, Wetton, & Hill 2013; Yamani et al., 2016). These studies suggest that, given sufficient impetus to seek and/or provide training (for example, by requiring new drivers to pass a hazard perception test), performance in this higher-order skill can be improved. One caveat to this conclusion was pointed out by McDonnald et al., (2015) who noted that no studies of HP training interventions have, to date, followed-up with the participants to identify whether the training had an impact on subsequent crash propensity. Since McDonald et al.'s assessment of the field, promising results have been found with a training intervention undertaken by the US National Highway Traffic Safety Adminstration, based on Don Fisher's Risk Awareness and Perceptual Training programme (Thomas, Rilea, Blomberg, Peck, & Korbelak, 2016). They found their brief hazard training intervention to reduce future collisions by over 23% in their male cohort, though the training was not successful with females.

Certainly, the introduction of the national HP test in the UK has ostensibly decreased road collisions. Wells et al., (2008) reported a 17.4% reduction in non-low speed collisions (where blame could be attached) linked to the introduction of the test in 2002. This result demonstrates the significant impact that the HP test has had on UK road safety, and raises the possibility that this could be of equal use to other countries who are facing even greater road safety challenges.

An international perspective on hazard perception

Although the UK was the first country to include an HP test as part of the licensing procedure, both Australia and the Netherlands have since developed their own HP tests. In addition, there are research groups around the world who have developed HP tests in their own countries, including Australia Spain, Germany, The Netherlands, Israel, Singapore, Malaysia, Canada, Hong Kong, China, Japan, and New Zealand (e.g. Borowsky et al., 2010; Cheng, Ng and Lee, 2011; Cocron et al., 2014; Gau, Yu and Hou, 2015; Horswill, Anstey, Hatherley, and Wood, 2010; Horswill, Hill and Wetton, 2015; Malone and Brünken, 2016; Isler, Starkey and Williamson, 2008; Lim, Sheppard and Crundall, 2013; Rosenbloom, Perlman, and Pereg, 2011; Scialfa, Pereverseff, and Bokenhagen, 2014; Shimazaki, Ito, Fujii, and Ishida, 2017; Ventsislavova et al., 2016; Vlakveld, 2011, 2014; Wang, Peng, Liang, Zhang, and Wu, 2007; Wetton, Hill and Horswill, 2011; Wetton, Horswill, Hatherley, Wood, Pachana, and Anstey, 2010; Yeung and Wong, 2015). Unfortunately, the results of many studies from around the world are mixed, with some researchers finding safe to perform better than les-safe drivers (e.g.Wallace and Horswill, 2007; Horswill, et al., 2015), while others fail to find this basic effect (e.g. Sagberg and Bjørnskau, 2006; Lim et al., 2013; Yeung and Wong, 2015).

It is difficult to pinpoint the reason why some studies successfully discriminate between safe and less-safe drivers, while others do not, as the precise design of these various tests can differ on many crucial points. The most interesting difference between these studies is the country in which they are conducted. Both the stimuli (the video clips containing the hazards), and the participants, are culturally specific to the region. There are wide cultural differences in the nature of driving, including both the legal and social rules that govern acceptable behaviour, which in turn influence the nature of the hazards. It is possible that some types of hazard are more prevalent in particular countries, and that some of these hazards may be less successful in differentiating between driver groups (Crundall et al., 2012; Crundall 2016), or are simply unsuitable for a hazard perception test. For instance, when one of the current researchers was filming hazard footage in Malaysia (Lim et al., 2013) many of the naturally occurring hazards did not make the final cut. The majority of these rejected hazards were interactions between the film car and motorcycles, which would overtake without warning and cut in front of the film car, necessitating urgent braking in some instances. The immediate appearance in the camera view of these motorcycles, would not have provided the more experienced drivers in the study with any precursors (i.e. visual clues) to help them predict the occurrence of the hazard (Pradhan and Crundall, 2017), and thus would be unlikely to find a performance difference between safe and less-safe drivers.

This touches on several other reasons why differences in findings might arise between research studies: there is no accepted standard for what constitutes a hazard, or how these clips should be edited and then presented, or even what response should be collected from participants (Ventsislavova and Crundall, 2018). Many research teams adopt an individual approach to developing hazard perception tests, making it difficult to compare studies across different countries when they have employed different methodologies and used different sets of clips. In fairness we should note that this is not necessarily a problem just across countries, as there are several studies conducted within the UK (again using different hazard clips) that have failed to replicate the basic behavioural differences between experienced and novice drivers (e.g. Crundall et al., 1999, Underwood, Ngai and Underwood, 2013).

To our knowledge, only Lim, Sheppard and Crundall (2013; 2014) have measured performance on the exact same test across two different countries[1]. In 2013 they compared Malaysian and UK drivers' hazard perception performance on clips filmed in both countries. They found that the UK drivers responded to many more hazards than the Malaysian drivers, especially when they were presented with Malaysian clips. A difference between novice and experienced drivers did not materialise however (in both Malaysian and UK participants). The authors suggested that cultural differences in hazard criterion (the internal threshold at

[1] Wetton et al., (2010) found novice/experienced driver differences with Australian participants viewing a UK test, but did not test UK participants. Still the positive result suggests some generalisability between two countries, albeit countries that are highly similar in terms of culture and road laws

which one considers an event to be a ‘hazard’) impacted more on test performance than experience. As Malaysian drivers typically encounter more hazards on the road than UK drivers, these events become normalised to the extent that a scenario must be extremely dangerous before they consider it to be a ‘hazard’, rather than just an everyday event. However, without finding a difference between UK novice and experienced drivers, they could not firmly conclude that the hazard perception test could not transfer between countries (as they could not establish the effect in the UK in the first place with their clips).

In the 2014 study they had more success. Using the same clips, they created a hazard prediction test. This test was created following Jackson et al.’s guidelines (2009): clips are suddenly occluded just as the hazard begins to materialise, and participants are asked ‘what happens next?’. In a modification to the free-response answers given by Jackson et al.’s participants, Lim et al. (2014) provided participants with 4 multiple-choice options from which to choose.

The rationale behind the hazard prediction test is that it isolates the predictive element of the hazard perception process (Pradhan and Crundall, 2017), providing a measure that records accuracy (unlike the traditional HP test), which is unconfounded by criterion bias (i.e. the participant’s response is not dependent on an internal threshold for reporting hazards, as a response time measure is; Crundall, 2016). Judging ‘what happens next?’ is independent of whether one thinks it poses a threat beyond your self-perceived level of driving skill. Following Jackson et al. (2009), several studies have demonstrated that this prediction test can discriminate between novice and experienced drivers (Castro et al., 2014, 2016, Crundall, 2016, Ventsislavova et al., 2016).

When Lim et al. (2014) presented the hazard prediction test to both Malaysian and UK participants, they found that it discriminated between novice and experienced drivers, regardless of the nationality of the participants, though the effect was only apparent with those clips that were filmed in the UK. A number of points are worthy of note here. First, this was the first study to use hazard prediction clips that had previously been used as a hazard perception test. The fact that the clips did not identify differences between experience groups as a hazard perception test, but did produce a difference between groups as a hazard prediction test, suggests that the latter approach is more robust. Secondly, the fact that the UK clips could discriminate between novice and experienced Malaysian drivers argues for some degree of cross-cultural generalisability. While the results of the Lim et al. studies (2013, 2014) are not completely clear cut, the data appear to favour the prediction test over the hazard perception test as a potential road safety export, though the two studies were never directly compared.

The current experiments

The current paper describes two studies that set out to assess whether two variants of the hazard perception methodology could successfully discriminate between experienced and inexperienced drivers across three countries (China, Spain and the UK), paving the way for the design and export of a culturally-agnostic test. While most studies of HP performance across countries use different stimuli and different test formats, the current studies used the same clips and identical methodologies, across a cohort of participants recruited in the three countries. All participants saw three sets of clips, with one set filmed in the UK, one set filmed in Spain, and a third set of clips filmed in China (e.g. all UK participants saw clips from China, Spain and the UK, etc.).

The first study compares participants' performance across countries (both in terms of participant nationality and clip origin) using a traditional hazard perception methodology which requires a timed button response to the appearance of a hazard. The second study recruited a new cohort of participants from across the three countries, and presented them with the same clips, but within a hazard prediction paradigm (i.e. the original hazard perception clips were edited to occlude just as the hazards onset, and participants were asked 'what happens next?'). We predicted that both tests would show differences between experienced and inexperienced drivers across the countries, though we were aware that the slim evidence that exists (Lim et al., 2013, 2014) suggests that the latter test might be more successful than the former.

Experiment 1

The first experiment compared UK, Spanish and Chinese participants' hazard perception performance for detecting hazards in three sets of clips filmed from each country. These three selected countries have very different cultures and traffic collision statistics. The World Health Organisation (2015) estimates the road fatalities of China, Spain and the UK to be 18.8, 3.7 and 2.9 deaths per 100,000 population, respectively, with an estimated 260,000 annual fatalities in China (though the officially reported number was just over 58,000 for 2013). Differences between officially reported statistics and WHO estimates reflect a number of measurement difficulties, such as trying to equate different definitions of a road collision fatality across different countries. For instance, in the UK an individual must die with 30 days of a collision to be counted, whereas in China the deadline for inclusion is 7 days.

While the safety records of the UK and Spain are much more comparable, they still differ markedly in terms of culture and road laws (with the most considerable difference being the side of the road on which they drive). Thus across all three countries we have a range of cultures, laws, and risk of collision, providing a demanding assessment for a culturally-agnostic hazard perception test.

Key to this study was the requirement that the country-specific tests were as similar as possible in all other ways. Thus, all clips from each country were filmed and edited for this specific study, rather than co-opting previously captured video footage for inclusion. From the experience of filming in Malaysia, thought was given as to how best capture hazards that might not be suitable for the single-camera forward view favoured in the official UK and Australian tests. In order to accommodate the potential increase in overtaking hazards that may occur outside the UK, we used additional cameras attached to the film vehicle to record the views that one would see in the rear view mirror, and the two side mirrors. These video streams were then synchronised with the forward view and edited into mirror placeholders created in a graphic overlay of a car interior. Mirror information has been used previously in hazard perception clips. For instance, Borowsky, Oron-Gilad, Meir, and Parmet (2012) included an inset rear-view mirror in their clips, though their clips did not require attention to the mirror information. Crundall, Crundall, Clarke and Shahar (2012), Shahar, Alberti, Clarke and Crundall (2010), and Shahar, van Loon, Clarke and Crundall (2012) included both side mirror and rear-view mirror information, inset into the forward view, in their hazard perception clips, though only the latter study required participants to use the mirror information to decide when it was safe to change lanes. The current study however combines mirror information with a graphic overlay to create a more immersive environment, providing precursors for hazards that appear from behind the film-car. We predicted that this test format would differentiate between experienced and inexperienced drivers in each country (using experience as a surrogate for crash likelihood). It was also considered likely that experience might interact with clip origin and participant nationality, such that UK experienced drivers may only out-perform UK inexperienced drivers on UK clips, etc. Such findings would at least demonstrate that the test format is culturally agnostic, if not the actual stimuli filmed in the three countries.

Method

Participants

One hundred and fifty three participants were recruited for Experiment 1. The sample was composed of drivers from three different countries (Chinese participants = 50, Spanish participants = 51, UK participants = 52). All of the drivers held full or provisional licences from their respective countries. Participants were split into experienced and inexperienced driver groups (46% experienced drivers and 54.05% inexperienced drivers). According to the literature, novice drivers are overrepresented in crashes in the first 12 months after licensure in comparison to experienced drivers (Foss et al., 2011, McCartt et al., 2003; Williams and Tefft, 2014; Pradhan and Crundall, 2017). Thus, for this study the experienced groups were defined in the following way: Drivers were considered 'experienced' if they had passed their driving test at least 1 year before the study, and had driven at least 600 miles (965 km) in the previous year (to ensure that our experienced participants were still active drivers). Inexperienced drivers included learner drivers (34%), those who had passed their test in the same year of the study, plus a small number of drivers (2%) who had passed in the last few years but reported very little exposure (<600 miles in the previous year). These classifications resulted in 19 experienced and 31 inexperienced Chinese drivers, 26 experienced and 25 inexperienced Spanish drivers, and 23 experienced and 24 inexperienced UK drivers. Due to low absolute numbers of reported collisions (4 Chinese, 5 Spanish and 5 UK drivers reported collisions in the past 12 months), these data were not used to define the groups.

Demographic details for each group can be found in Table 1. Over all three countries the average experienced driver was 29.5 years old, passed the driving test in 2005 (with 10 years of experience), and drove 11804 miles per year. The average inexperienced driver was 21.1 years old, passed the driving test in 2014 and had an annual mileage of only 63 miles.

Participants from the three countries were recruited either from the respective Universities involved (Granada, Nottingham Trent and Tsinghua Universities), and from local driving schools. All of the participants were unpaid volunteers.

Demographics	Chinese Participants		Spanish Participants		UK Participants	
	Novice	Exp'd	Novice	Exp'd	Novice	Exp'd
Study 1: Hazard Perception						
Total N (female N)	31 (13)	19 (2)	25 (17)	26 (8)	24 (13)	23 (21)
Age	22	28.6	19.2	35.9	22.1	24.1
Post-licence Experience (years)	1	6	1	17	1	7
Annual Mileage	82.8	4274.3	28	22041.9	78.2	9095.7
Study 2: Hazard Prediction						
Total N (female N)	24 (9)	26 (4)	25 (21)	27 (4)	27 (18)	23 (23)
Age	22.7	25.3	20.2	40.9	19.4	24.4
Post-licence Experience (years)	1	5	1	21	1	7
Annual Mileage	33.7	5474	28.9	20183	266.7	5587

Table 1 Mean demographic values for participants in study 1 and 2.

Materials and apparatus

To create the hazard perception stimuli filming was undertaken in China, Spain and the UK. The forward view was recorded with a mini HD video camera attached to the inside of the front windscreen via a suction mount. The rear-view mirror footage was recorded via a camera attached to the inside of the rear window via suction mount. Two additional cameras were attached externally to the passenger and driver side windows, pointing behind the car to capture side-mirror footage. These cameras were also fixed via suction mounts and they were tethered to the car for safety. The driver of the film car in each country was an experienced, native driver, with previous experience of conducting driving-safety research. Filming took place across a variety of times, but always in daylight and with clear weather conditions. The filmed environment in each country included city driving (Beijing, Granada, Nottingham), suburbs, and rural locations. Ten clips were chosen from each country to create the hazard perception test (with 30 clips selected in total). Clips varied in length from 31s to 64s and each clip included one *a priori* hazard identified by our team of transport researchers from across the countries. All hazards were captured naturalistically. In addition to the actual hazard (see Table 2 for a description of the individual hazards), these clips typically included several other potential hazard sources (i.e. precursors that did not develop into hazards, Pradhan and Crundall, 2017). Hazards were defined as events where an object, either individually or in confluence with other objects, becomes set on a trajectory that would lead to a collision without corrective action undertaken by either the object or the driver of the film car. For example, a pedestrian on a sidewalk is considered a precursor to the hazard. When the pedestrian steps into the road however, we consider that she has become a hazard. This is termed the hazard onset and marks the start of the scoring window. As soon as a counteraction is instigated to avoid the hazard (the pedestrian may jump back on the sidewalk, or the driver of the film car may brake or swerve to avoid a collision), this is considered the hazard offset and the scoring window closes.

Once the clips had been selected, the mirror footage was synchronised with the forward view and edited into mirror placeholders that were contained in a graphic overlay of the interior of a car. The graphic overlay was generated from internal photographs of a Ford Focus. The A-pillars and roof were designed to be semi-transparent, allowing the forward view to be seen through these sections of the overlay, although at a reduced fidelity. This was done to simulate the fact that real-world obscuration by A-pillars is offset somewhat by stereopsis and small head movements. Equally, head movements can often bring objects back into the visual field that are obscured by the roof (e.g. If one is first in a queue of vehicles at a red traffic signal, the nearest set of signals can easily be hidden by the roof, necessitating the driver to lean forward slightly to be able to look up at the red light). The dashboard, and mirror placeholders were however fully opaque. The final edited video clips created a seamless driving experience. When passing a car travelling in the opposite direction the vehicle would disappear briefly (into the driver's blind spot) before reappearing in the mirrors. Screen shots from each country can be viewed in Figure 1.

In order to ensure comparability of instructions across the three countries, the UK instructions were subjected to a Chinese and Spanish forward-backward translation (following the guidelines of International Test Commission; ITC, 2010). This was undertaken to ensure that the participants understood what was meant by a "hazardous situation" and how they should respond. The translation into Chinese and Spanish was performed by a team consisting of three bilingual experts with a high level of expertise in Chinese and Spanish culture, traffic regulations and driving habits.

Clips were displayed on a Lenovo (ThinkPad) computer with resolution of 1920x1080 and screen size of 34.5cm x 19.5cm in all three countries and the programme used was E-Prime 2.0 Software (Psychology Software Tools, 2012). Participants responded with a mouse connected to the laptop.

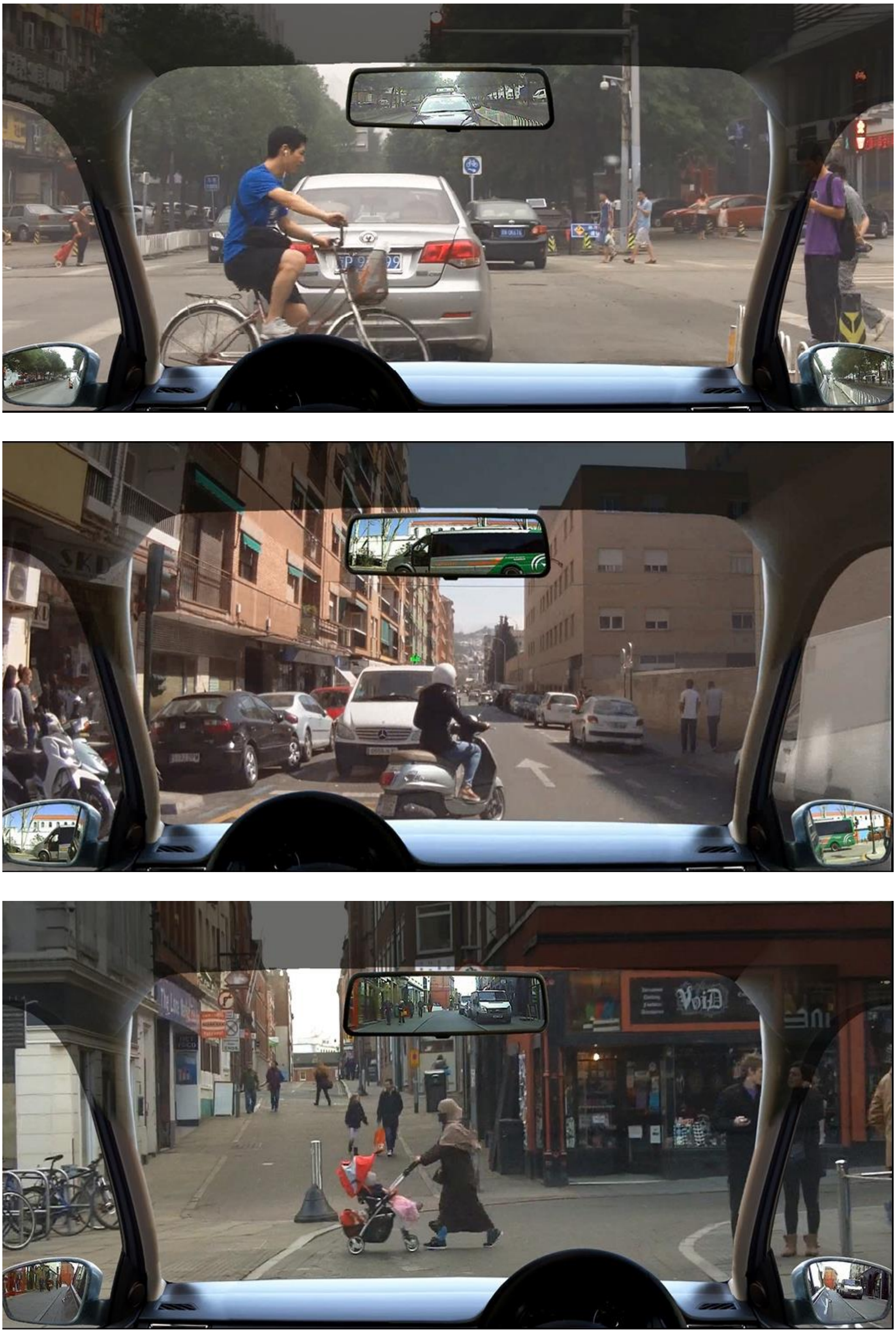


Figure 1. Three screen shots taken from hazard perception clips filmed in China (top panel), Spain (middle panel) and the UK (bottom panel).

Clip Number	Hazards (*with occlusion points for experiment 2 italicised*)	Duration of the clip (ms)
	CHINESE CLIPS	
1	A pedestrian is visible at the right edge of the road, looking to cross. The pedestrian is obscured by a turning car at the point of stepping into the road. By the time the pedestrian is visible again, he has already stepped into the road becoming a hazard. *For experiment 2, the clip occludes just as the pedestrian starts to become visible as the obscuring car moves past.*	55000
2	There are parked cars on both sides of a narrow street that might occlude pedestrians. A pedestrian steps into the road in front of you, from between two parked vehicles. *The clip occludes as the pedestrian first becomes visible stepping out from between the parked cars.*	57000
3	A gap in a long line of parked vehicles on the left side of a one-way street indicates the presence of a side road. A cyclist emerges from the side road, obscured by the parked vehicles, and makes a wide turn in front of your vehicle, before cycling towards you. *The clip occludes as the front wheel of the bicycle enters the view.*	26000
4	Your car slows on approach to a junction. A cyclist approaches from the left and is partially obscured by the A-frame of the semi-transparent graphic overlay. The cyclist cuts directly	63000

	across your path. *The clip occludes as the cyclist makes a change in direction to cut across your path.*	
5	Your car is driving slowly and there are parked cars on the right. A parked car on the right side of your lane indicates late before attempting to pull out in front of you. *The clip occludes following one flash of the indicator from the manoeuvring car.*	57000
6	A car immediately behind you, visible in the rear-view mirror and left side mirror, decides to overtake by entering a slip road to your left. It is forced to immediately pull back into your lane, in-front of you, as the slip road ends. *The clip occludes when the car is no longer visible in the left mirror, but flash of it is visible in the left window.*	31000
7	A car behind you is visible in the rear-view mirror and left side mirror. The car undertakes you on the left by entering a bus lane. Once past you, it cuts into your lane and is forced to brake suddenly due to slowing traffic ahead. *The clip occludes when the car is no longer visible in the right mirror, but it quickly appears next to the right window of your car*.	50000
8	A car behind you is visible in the rear-view mirror and left side mirror. While attempting to exit a multilane road, the car from behind accelerates to undertake your vehicle, forcing you to hold off moving into the desired lane. *The clip occludes when the car is no longer visible in the right mirror.*	62000
9	A lorry approaches fast from the right-hand side. The lorry enters the main road from a side road on the right, cutting into	38000

	your lane. *The clip occludes at the moment in which the lorry is about to enter into the main road.*	
10	A car behind you is visible in the rear-view mirror and left side mirror. The car indicates with the front lights. Then undertakes via a bus lane at speed, immediately cutting in front of your vehicle and braking. *The clip occludes when the car is no longer visible in the right mirror.*	55000
	SPANISH CLIPS	
1	A gap in a long line of parked vehicles on the left side of a one-way street indicates the presence of a side road. A motorcyclist emerges from the side road, partially obscured by the parked vehicles, and enters the main carriageway immediately in front of your vehicle. *The clip occludes when the front part of the motor is visible.*	41000
2	A pedestrian is stood in the road next to a parked car waiting for her friend to exit the car. As your car approaches the driver's door of the parked vehicle opens. *The clip occludes when the pedestrians stops next to the car.*	42000
3	On entering a side road, a rider on a scooter is checking over her shoulder in order to pull out around a vehicle blocking her lane. She then pulls out in front of your vehicle, as you finish turning into the side road. *The clip occludes following one flash of the indicator of the scooter.*	34000
4	While travelling on a dual carriageway, in the distance a pedestrian enters from the left side of the road. The pedestrian	37000

	continues to cross the street, forcing you to slow and stop. *The clip occludes at the moment when the pedestrian enters the road.*	
5	A car emerges from a side road and stops in front of you before indicating that it is going to reverse into a parking space at the road edge. *The clip occludes following one flash of the indicator from the car.*	53000
6	A van is approaching fast from a side road on the right. The van tries to pull out and it halts abruptly when already partially out of the road, but nevertheless forces you to brake suddenly. *The clip occludes when the van is approaching from the right and almost enters your lane.*	48000
7	A car ahead stops on a zebra crossing due to congestion ahead. A pedestrian, unable to cross on the actually crossing, steps into the road slightly in advance of the zebra crossing. As she steps out she is partially obscured by parked vehicles on the right. *The clip occludes just as the pedestrian first become visible.*	42000
8	A double-length (‘bendy’) bus in the right lane indicates and pulls off from a bus stop immediately in front of you, after you have just exited from a roundabout. *The clip occludes when the bus turns to enter the road and following a flash of the indicator of the bus.*	50000
9	While driving on a dual carriageway, a motorcycle undertakes in the right lane and is forced to pull in-front of your car as traffic in the right lane slows due to congestion. *The clip*	26000

	occludes when the motorcycle is no longer visible in the right mirror, but part of it is visible at the right window.	
10	A pedestrian is approaching from the left, partially obscured by a pillar. The pedestrian crosses the road ahead from the left. *The clip occludes when the pedestrian is approaching the zebra crossing.*	47000
	UK CLIPS	
1	A car ahead overshoots a red traffic signal. As the cross traffic begins to enter the junction, the reversing light of the car turns on and the car ahead reverses towards you. *The clip occludes following initial illumination of the reversing light.*	64000
2	You are driving in a street with shops and parked vehicles on the left side. There is a pedestrian coming out of one of the shops, approaching the street. The pedestrian steps out from between two parked cars on the left just as you accelerate after waiting in standing traffic. *The clip occludes when the pedestrian turn his head to look at your car*.	49000
3	A distracted pedestrian is walking towards the street. The pedestrian crosses the road from the right without looking. *The clip occludes when the pedestrian approaches the road in order to cross*.	54000
4	A car behind you is visible in the rear-view mirror and left side mirror. The car from behind undertakes you on the right on a multilane road. Once past you, it cuts into your lane and is forced to brake suddenly due to a red traffic light. *The clip*	43000

	occludes when the car is no longer visible at the left mirror, but a flash of it is visible on the left window.	
5	There are pedestrians on both side of a narrow street. A pedestrian with a child’s push chair enters the road from the right without looking. Her entrance is partially obscured by pedestrians standing on the right. *The clip occludes when the push chair is partially visible among the pedestrians.*	32000
6	A bus in a bus lane signals to pull away from a bus stop. Due to parked vehicles ahead in the bus lane, it pulls out into your lane forcing you to stop. *The clip occludes following one flash of the indicator from the bus*.	31000
7	A car ahead emerges from a side road on the left and crosses your lane (too far away to be considered a hazard). It then indicates again and immediately cuts across your lane once more to park in a layby. This second lane crossing is close enough to your vehicle to constitute a hazard. *The clip occludes just before the car starts to cross your lane and following one flash of the indicator of the car*.	42000
8	A car behind you is visible in the rear-view mirror and left side mirror. The car from behind overtakes your vehicle on a blind rural bend. The appearance of an oncoming vehicle in the opposite lane forces the overtaking vehicle to pull back into your lane immediately in front of you. *The clip occludes when the car is no longer visible at the rear-view mirror and the right mirror*.	34000

9	While travelling at speed along a country road, a blind bend ahead reveals a queue of standing traffic, forcing you to slow and stop. *The clip occludes immediately after passing the blind bend, when the brake lights of the cars ahead are partially visible.*	70000
10	While your car is slowing due to congestion ahead, a pedestrian looks over his shoulder before deciding to run in front of your vehicle forcing to slow more abruptly than otherwise required. *The clip occludes at the moment when the pedestrian is visible at the left pillar and is looking at your car.*	35000

Table 2. A description of the a priori hazards selected within each clip

Design

A 2x3x3 mixed factorial design was used. The between-group factors were the driving experience of participants (experienced vs. inexperienced) and their nationality (Chinese vs. Spanish vs. UK). The within-group factor was the clip origin (China vs. Spain vs. UK). The dependent variables included the percentage of hazards that participants correctly identified and their response times to these hazards.

Correct identification of a hazard was defined as a button response that fell within a temporal scoring window for each hazard. This scoring window began at hazard onset and terminated at hazard offset. Onset was defined as the point where a hazard begins to develop and will eventually pose a threat (e.g. a car ahead begins to edge out of a line of standing traffic in front of the film car; a pedestrian steps off the sidewalk, etc.). Offset is defined as the point at which the hazard was no longer a threat (e.g. corrective action had been taken by one of the road users to avoid a collision). The average scoring window was 5350 milliseconds from onset to offset.

At the end of each clip, participants were also required to rate each clip for the level of hazardousness presented on a Likert scale from 1-7, with higher numbers reflecting increasing levels of danger (with ‘not at all hazardous’ to ‘extremely hazardous’ as the anchors). Clips from the three countries were presented in three different blocks (10 clips per country). Both the order of the clips within the blocks, and the order of the blocks were randomised for each participant. The design of the experiment was approved by the College of Business, Law and Social Sciences Research Ethics Committee at Nottingham Trent University, UK.

Procedure

Participants were seated in front of the screen and viewed the on-screen instructions in their native language. They were asked to fill in a demographic questionnaire which included information such as age, sex, year of obtaining their driving license, driving collisions in the past 12 months, and miles/kilometres driven in the past 12 months. Participants were seated 60 cm from the screen. The screen measured 34.5cm x 19.5cm. When participants were sat at a distance of 60 cm the screen subtended 26.27 degrees along the horizontal axis, and 14.91 degrees in the vertical axis.

Participants were told that they would see 30 video clips from the driver’s perspective, recorded in three different countries, and that each contained at least one hazardous situation. They were asked to view these clips as if they were the driver, and to press the mouse button as soon as they saw a hazard occurring. A hazard was defined as an object or event in the road environment that could increase the risk of a collision if an evasive manoeuvre such as braking or steering was not performed (following Crundall, 2016). It was made clear that participants did not need to locate the hazard on the screen using the mouse, but merely had to press the button to record their response. After each clip they were asked to rate how hazardous they thought that particular situation was on a scale from 1 to 7 by pressing the corresponding button on the keyboard (1 = not at all hazardous; 7 = extremely hazardous). Before the start of the experiment each participant saw a practice clip from their own country in order to familiarise themselves with the task. If the participant failed to perform on the practice task as expected, the experimenter explained the instructions again. In total the experiment took an average of 35 minutes.

Results

For this experiment 5 of the 153 participants were removed (N=148): four of them due to excessive clicking (>60 clicks per block of videos) and one of them for being an inexperienced driver with high mileage (i.e. >6000 miles in their first year of driving). These five outliers were all UK drivers.

There were two main measures of interest: the proportion of hazards correctly identified (i.e. where a response was made within the temporal scoring window), and the response times associated with these mouse clicks. Traditionally this type of factorial design is analysed using mixed ANOVA, but such an analysis has a number of shortcomings. First, it treats stimuli (clips) as a fixed factor rather than a random factor. The current design has two fully-crossed random factors (participants and clips). Ignoring the second random factor can inflate Type I error rates (Judd, Westfall & Kenny, 2012). Perhaps more importantly, treating clips and participants as random effects increases our ability to generalize beyond the sample of clips used in the experiment. Second, it treats discrete outcomes, such as correctly detecting a hazard, as continuous. A more appropriate model is therefore a multilevel generalized linear model, that allows a discrete response to be modelled with fully-crossed random factors (Baguley, 2012). Such models were previously difficult to fit, but recent software developments have made the process easier. We have used the free, open source environment R (R Core Team, 2018), with the package lme4 (Bates, Maechler, Bolker & Walker, 2015). In all cases we fitted a sequence of models starting with intercept and random effects only, adding all main effects and then adding higher order interactions (starting with two-way interactions). Effects were tested using likelihood ratio tests to compare a model with all effects of the same order (e.g., two-way interactions) to a model where the effect of interest is dropped.

Accuracy of hazard detection (based on whether participants responded within the scoring window, coded 1 or 0) and number of clicks (as counts) were analysed to compare performance across driver experience (experienced vs. novice drivers), participant nationality (Chinese, Spanish, or from the UK), and the origin of the clips (China, Spain or the UK). Between-group effects and within-group effects were further explored with 95% posterior probability intervals comparing Chinese Drivers to Spanish drivers, and Spanish drivers to UK drivers. Here, and in subsequent analyses, descriptive statistics associated with each analysis are estimates derived from the model rather than the raw data (although in all cases the differences are very small). The chief difference is that model derived estimates exhibit shrinkage towards more typical units – meaning that they are less influenced by unusual or extreme participants or clips.

Response accuracy to hazards

Participants were considered to have correctly responded to a hazard if they pressed the mouse button within the hazard window for each specific clip (the mouse cursor was not visible on the screen and the location of the mouse was not important). Responses were analysed using multilevel logistic regression (with each data point modelled as a Bernouilli trial). An intercept only model (with no predictors) estimated the *SD* of the participant random effect as 0.84 and the *SD* of the clip random effect as 1.15 indicating that only 35% variation at level 2 of the model is attributable to participants – with variability in clips accounting for the majority (65%) of level 2 variance. This indicates that a traditional ANOVA analysis – that treats variation between clips as zero – would substantially underestimate standard errors. The deviance (likelihood ratio Chi Square, G^2) for the intercept only model was 4459.0 and decreased dramatically to 4432.8 for a model including main effects of nationality, experience and clip type. This improvement in model fit was statistically significant, ΔG^2 (5) = 26.1, $p < .0001$. In addition, G^2 decreased substantially for a model with all two-way interactions, ΔG^2 (8) = 12.6, $p = .12$, with a negligible improvement with the addition of the three-way interaction, ΔG^2 (4) = 5.8, $p = .21$. The main effects model and two-way model appear to be the most informative (balancing goodness of fit and the effective number of predictors).

The pattern of accuracy across all conditions is shown in Figure 2. There was no indication that experience impacted accuracy with novice drivers slightly, but non-significantly, more accurate on average, ΔG^2 (1) = 0.2, $p = .69$. Nor were main effects detected for clip origin, ΔG^2 (2) = 0.65, $p = .72$.

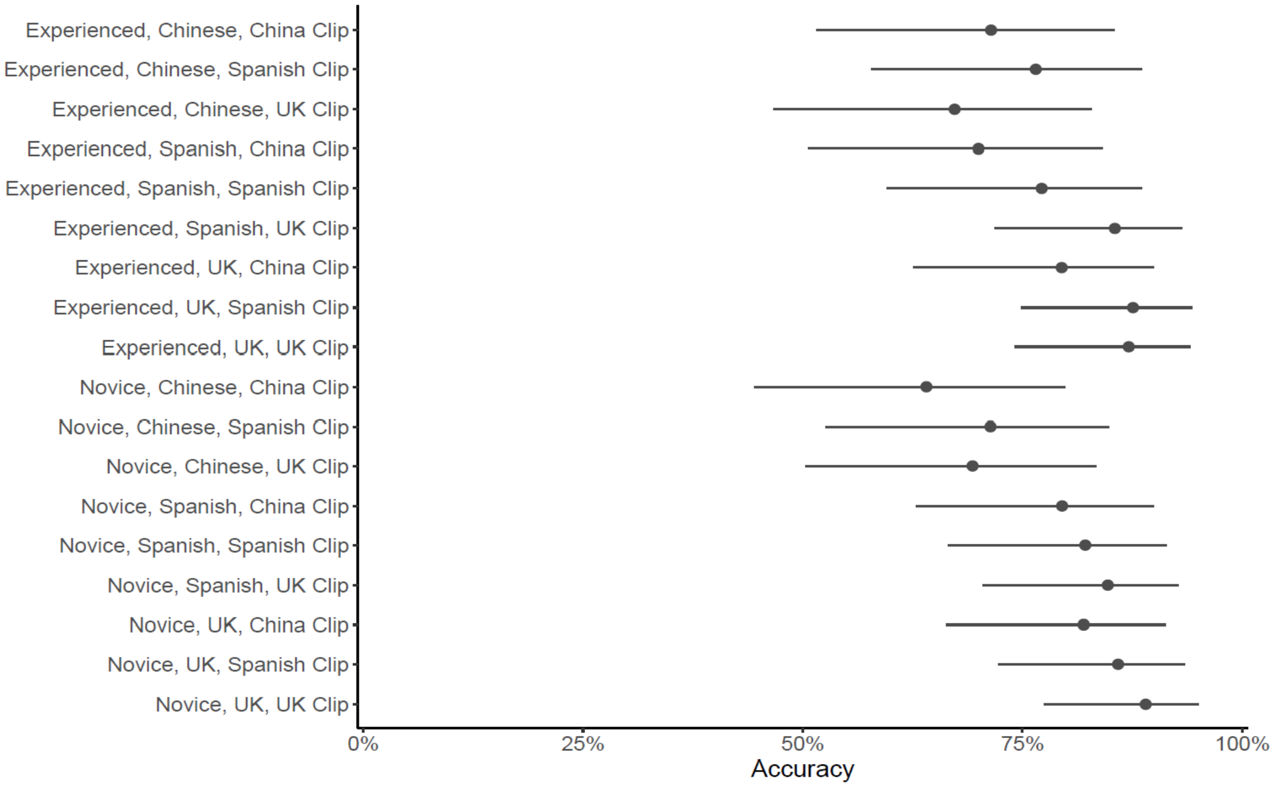


Figure 2. The percentage of hazards correctly responded to by all participant groups across the three countries of origin for the clips (with error bars).

However there was a main effect of participant nationality, ΔG^2 (2) = 25.5, p < .001. On average accuracy was 70.1%, 95% CI [59.3, 79.1], for Chinese drivers, 80.3%, 95% CI [71.6, 86.7], for Spanish drivers and 85.4% 95% CI [78.3, 90.5] for UK drivers. Follow-up tests indicate that differences between all three nationalities were statistically significant.[2] The odds ratio (*OR*) of the difference between Chinese and Spanish drivers was 0.58, p < .005, 95% CI [0.41, 0.81], whilst between Chinese and UK drivers it was, 0.40, p < .005, 95% CI [0.28, 0.57], narrowing to 0.69, p < .005, 95% CI [0.49, 0.99], between Spanish and UK drivers. Thus the data suggest a clear pattern of differences between drivers of different

[2] With only three means no correction is required for multiple testing. Type I error for the complete null hypothesis is protected by the initial likelihood ratio test and the number of Type I errors cannot exceed one for the three pairwise tests (Shaffer, 1986; Baguley, 2012).

nationalities – one that reflects the rank order of these countries in road safety according to the World Health Organisation; WHO, 2015).

A nationality by clip origin interaction was also detected, ΔG^2 (4) = 10.5, $p < .05$. This is depicted in Figure 3.

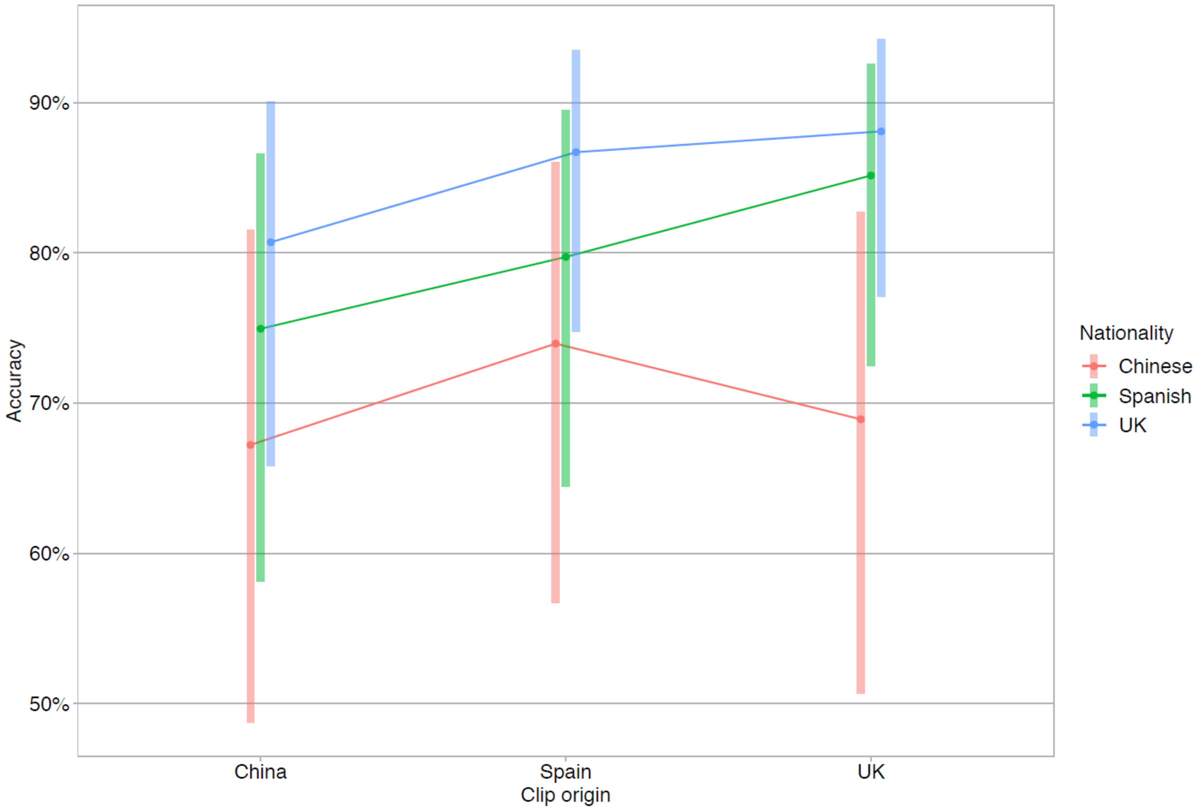


Figure 3. The percentage of hazards correctly responded for each block of clips across the three groups of participants (with error bars).

The pattern of accuracy indicated by the main effect showed that UK drivers were the most accurate on average, followed by Spanish drivers with Chinese drivers the least accurate. Additionally, the difference between Chinese and UK or Spanish drivers is particularly large for the UK clips. To confirm this we followed up the significant interaction with an interaction contrast comparing the difference between the UK and Spanish drivers and Chinese drivers for the UK and non-UK clips respectively. This contrast was statistically significant, G^2 (1) = 9.6, $p < .005$, explaining much of the deviance in the interaction and indicating that the Chinese drivers fared particularly poorly at detecting hazards in the UK clips relative to UK or Spanish drivers.

Response times to hazards

In order to calculate response times (RTs) for the hazards, hazard onsets and offsets were defined for each clip. Hazard onset times for each clip were subtracted from button-press times to give the RTs. Where participants failed to make a response during a particular clip, they were assigned a maximum response time plus 1 millisecond. The resulting data were therefore right-censored and, as is common with response times, positively skewed. To address these features of the data a multilevel regression model treating clips and participants as random factors and modelling response times as a right-censored lognormal distribution was employed. This is superior to standard approaches to such data that treat censored data as missing or treat censored responses as known with certainty. A 2x3x3 factorial model was fitted as a Bayesian model using the probabilistic programming language Stan (Carpenter et al., 2017) via the R package brms (Bürkner, 2017) using the weakly informative default priors. As brms does not provide frequentist likelihood ratio tests, models were compared using the information criteria WAIC (Vehtari, Gabry, Yao and Gelman, 2018) and followed up using 95% Bayesian central posterior probability intervals.

For an intercept only model (with only random effects) WAIC was 56559.7 decreasing to 56553.5 (ΔWAIC = 6.2) for a model with all main effects – indicating a substantial improvement in fit.[3] Figure 4 shows the predicted mean response times by

[3] WAIC, like over information criteria, is on the same scale as a likelihood chi-square statistic. To aid interpretation it can be helpful to scale this as a likelihood ratio (or, more accurately, Bayes factor). In this case a change in WAIC of 6.2 is equivalent to a Bayes factor of $e^{3.1}$ = 22.2 (with a posterior probability of .96 in favour of the main effects model).

condition. The two way model (WAIC = 56551.6) provided only a modest additional improvement in fit (ΔWAIC = 1.9). Accordingly only main effects were followed up further.

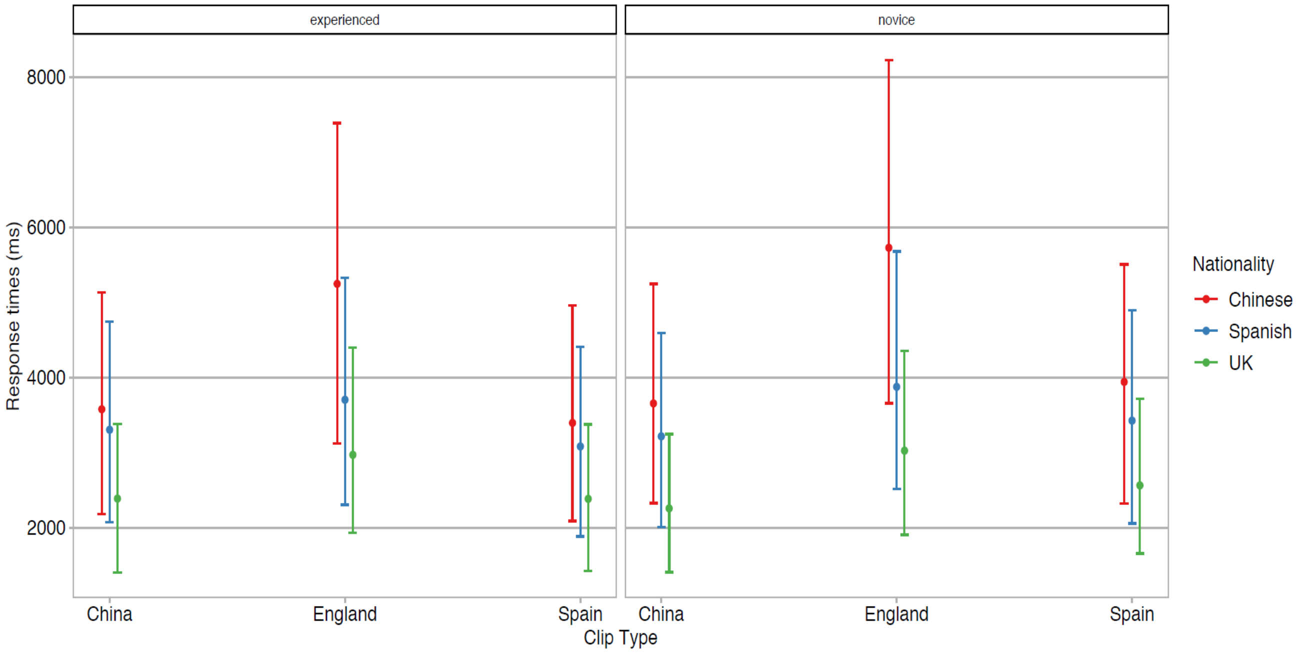


Figure 4. Response times across all participants nationality, experience groups and 3 clips sets recorded in each country (with error bars).

There was no substantial effect of experience on predicted mean response times with experienced drivers or for clip origin. However, there was evidence of differences between response times for nationalities. This pattern is clear from Figure 4 with UK drivers having the fastest mean predicted response times, *M* = 2254, 95% CI [1358, 3655], with typically slightly slower responses by Spanish drivers, *M* = 2974, 95% CI [1810, 4845], and the slowest for Chinese drivers, *M* = 3622, 95% CI [2209, 5835].

Extra hazard responses

In addition to the analysis of the two main DVs above, we also calculated the number of additional responses that participants made while watching the clips, above and beyond those responses that correctly identified the hazards. The *extra hazard response rate* is the number of mouse click responses made during an entire video that were not considered to be a correct response to the pre-defined hazard (and thus included responses that fell outside the hazard windows - potentially including responses to precursors - and also any responses in the hazard window beyond the initial response to the hazard). As the Chinese, Spanish and UK blocks varied in total duration (8 m 23 s, 7 m, and 7m 57 s, respectively) we modelled the responses as a Poisson count variable with an offset to account for the extra exposure for the duration of each clip (Baguley, 2012). The resulting multilevel generalized linear model, which included participant and clip as random factors, therefore estimated the extra hazard responses per minute (EHR/m) and was fitted using lme4. As with the accuracy analysis the intercept only model (G^2 (3) = 1355) was a worse fit than a model with all main effects (ΔG^2 (5) = 29.0, $p < .0001$), which in turn was a worse fit than a model with all two-way interactions (ΔG^2 (8) = 36.1, $p < .0001$). Adding the three-way interaction did not further improve the model (ΔG^2 (4) = 2.2, $p = .70$). The mean EHR/m by condition is shown in Figure 5. Main effects were found for nationality, G^2 (2) = 16.9, $p < .0001$, and clip origin, G^2 (2) = 11.5, $p < .005$, but not driver experience, G^2 (1) = 0.1, $p = .70$.

For nationality the rate of extra responses was higher for Spanish (EHR/m = 1.45) than Chinese drivers (EHR/m = 0.94) and the rate ratio (*RR*) for this difference was statistically significant, *RR* = 0.65, 95% CI [0.45, 0.77]. The rate for UK participants' extra responses (EHR/m = 1.59) was also higher than for Chinese participants. This difference was significant, *RR* = 0.65, 95% CI [0.50, 0.84], however rates for Spanish and UK participants were not significantly different, *RR* = 0.91, 95% CI [.70, 1.18].

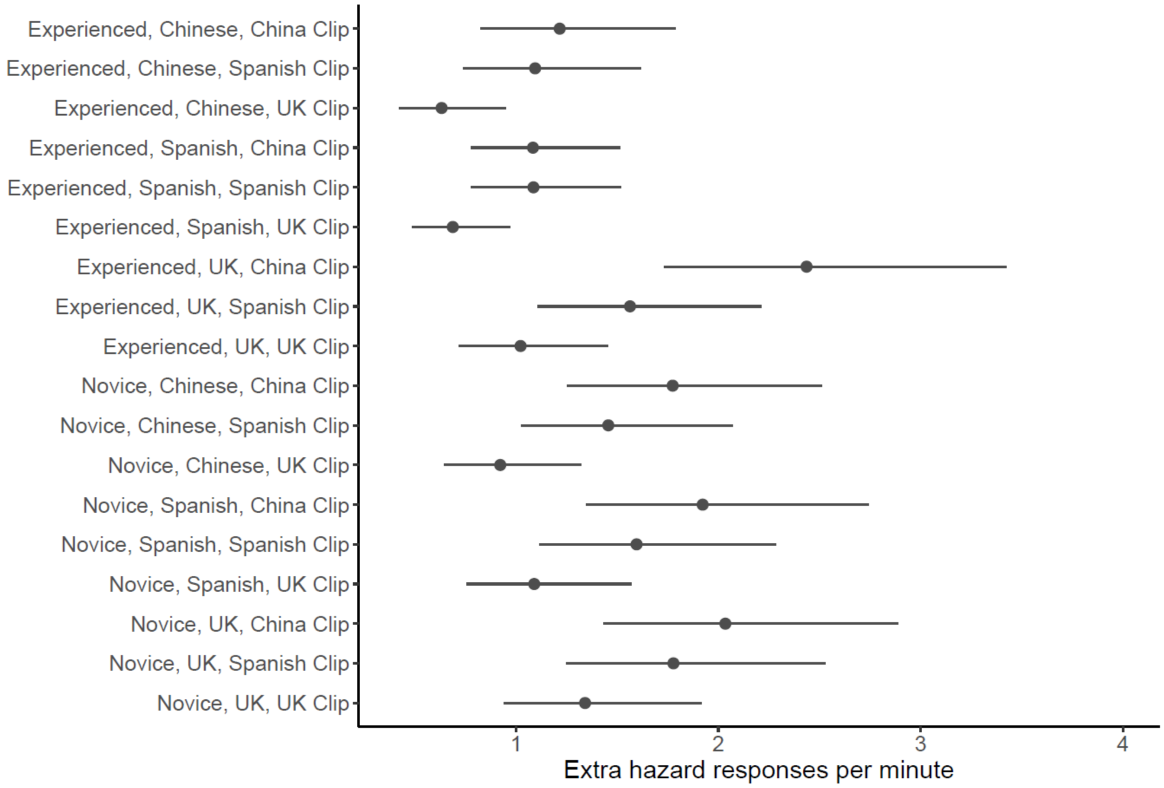
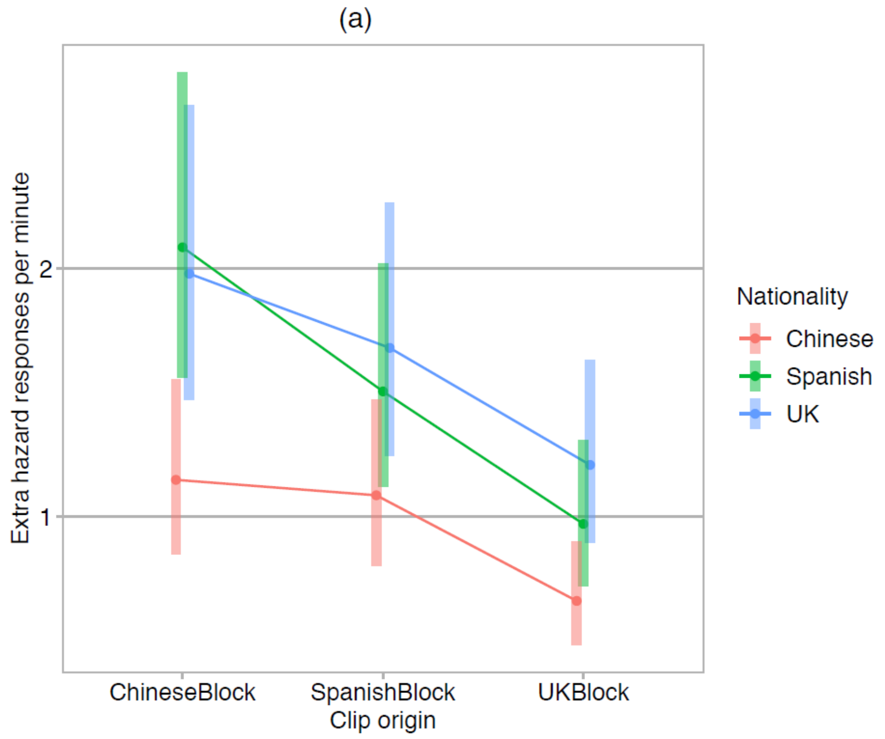


Figure 5. Extra hazard responses across all participant groups and 3 sets of clips (with error bars)

For clip origin the rate of extra responses was lower for UK (EHR/m = 0.92) than Chinese (EHR/m = 1.68) or Spanish clips (EHR/m = 1.40). The difference in rates was statistically significant for UK clips versus both Spanish clips, *RR* = 1.52, 95% CI [1.10, 2.11], and for the UK versus Chinese clips, *RR* = 1.83, 95% CI [1.32, 2.53]. However there was no significant difference in rates between the Chinese and Spanish clips, *RR* = 1.20, 95% CI [0.87, 1.66].

A significant interaction was found across clip origin and nationality, G^2 (4) = 21.2, $p < .0005$, and across clip origin and driver experience, G^2 (2) = 9.7, $p < .01$. However no experience by nationality interaction was detected, G^2 (2) = 1.3, $p = .53$. These interactions are plotted in Figure 6a and Figure 6b.



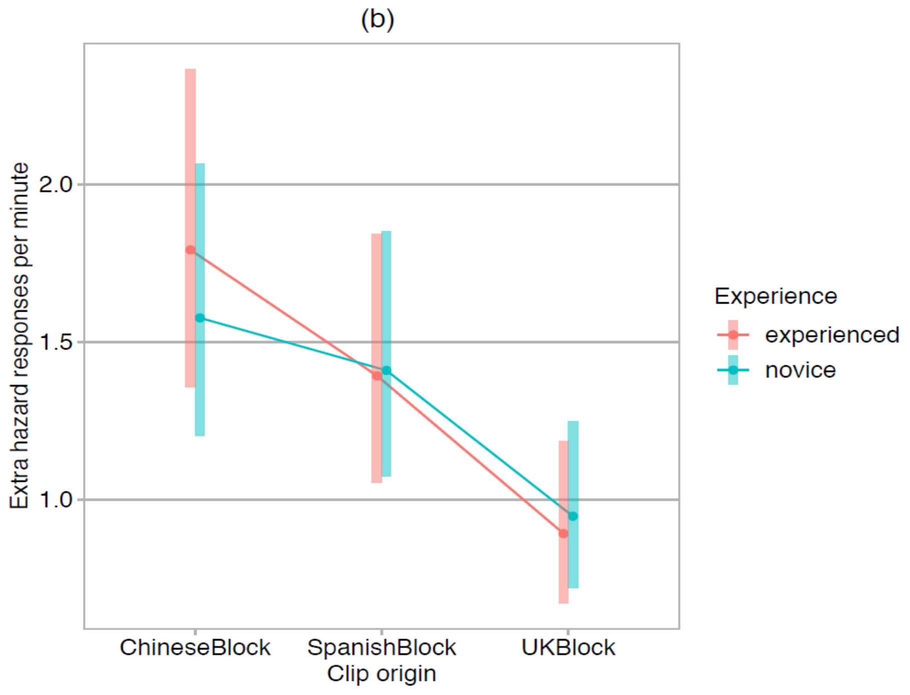


Figure 6a. Extra hazard responses across clip origin vs nationality (with error bars)

Figure 6b. Extra hazard responses across clip origin vs experience (with error bars)

Figure 6 helps clarify the patterns of extra hazard responses. Panel (a) indicates that the lower rate of hazard responses for Chinese participants relative to UK and Spanish participants is obtained across all clip types but is particularly pronounced for clips of Chinese origin. In contrast, panel (b) reveals that the lower hazard response rate for UK clips is observed regardless of experience, but that the difference between the Chinese and UK clips is larger for the experienced than inexperienced drivers. Simple main effects for the Chinese versus Spanish differences are not significant for either experienced or novice drivers (both $p > .05$).

A further analysis addressed whether extra hazard responses were related to accuracy by adding accuracy (0 or 1) as a covariate to the two-way model. Accuracy was positively, albeit modestly, associated with the rate of extra hazard responses, *RR* = 1.08, 95% CI [1.01, 1.15], G^2 (1) = 4.7, $p < .05$. This suggests that a successful hit rate appears related to the overall number of extra hazard responses that participants made, despite having already removed the 4 UK participants who were considered ‘excessive responders’ (more than 3 *SD*s above the average participant per block, i.e. more than 60.4 responses during the 10 clips from a particular country).

Hazardousness ratings for all hazard clips

Following each clip, participants were asked to provide a hazardousness rating on a scale of 1 to 7 (where higher numbers reflect greater levels of perceived hazardousness). These ratings were analysed with a three-way factorial design (participant experience vs. nationality vs. clip origin) using multilevel ordinal logistic regression via the ordinal package (Christensen, 2018). The intercept only model (G^2 (8) = 14226) was a worse fit than a model with all main effects (ΔG^2 (5) = 40.0, $p < .0001$), which in turn was a worse fit than a model with all two-way interactions (ΔG^2 (8) = 40.8, $p < .0001$). Adding the three-way interaction did not further improve the model (ΔG^2 (4) = 1.8, $p = .78$).

It was observed that Chinese clips were rated as more hazardous than Spanish Clips (4.5 vs. 4.14) and the Spanish clips were rated as more hazardous than the UK clips (4.14 vs. 3.84), but the main effect did not reach statistical significance, G^2 (2) = 5.5, $p = 0.06$. Post hoc tests with a Hochberg correction, which does not require significance of the main effect (e.g., see Baguley, 2012), detected only a difference between UK and Chinese clips (adjusted $p = .03$). Nationality, however, produced a significant main effect, G^2 (2) = 31.8, $p < 0.0001$, with Spanish drivers giving the highest ratings (4.73) followed by UK drivers (4.14) and Chinese drivers giving the lowest hazardousness ratings (3.60). These differences are statistically significant using Hochberg-corrected post hoc tests (adjusted $p < .02$ for all tests). The final factor of driver experience did not reveal a difference between the two groups (4.2 vs. 4.1 for experienced and inexperienced drivers, respectively), G^2 (1) = 1.4, $p = 0.23$.

An interaction was noted between clip origin and participant nationality , G^2 (4) = 35.5, $p < 0.001$. This appeared to reflect a slight difference in the Nationality pattern for different Clip types. Using Hochberg-corrected tests of simple main effects within each Clip type, the same pattern of differences between driver nationalities was found for all Clip types (all adjusted $p < .05$) except for the Spanish-UK difference for the Spanish clips (adjusted $p = .13$) and the Chinese-UK difference for the UK clips (adjusted $p = .06$). Thus it appears that the finding of higher ratings for Spanish drivers than UK drivers may not hold for Spanish clips and the lower ratings for Chinese drivers than UK drivers may not hold for UK clips.

Discussion

Experiment 1 revealed no differences between experienced and inexperienced drivers in regard to the two main dependent measures: response times to hazards, and the percentage of hazards correctly responded to. Thus it is hard to conclude that the hazard perception methodology is suitable for export to other countries when we cannot identify differences between any driver group on the basis of experience, regardless of nationality. If at least the UK clips could produce a difference between UK experienced and inexperienced drivers, then we could feel comfortable that the basic test replicated previous work in the field, but this was not the case. Admittedly, in order for this effect to have risen to our attention it would have had to evoke a three-way interaction between participant nationality, clip origin and driver experience. Failures to hit such high goals may always raise suspicions of a lack of statistical power, but even when UK experienced drivers' accuracy rates and response times are directly compared to those of UK inexperienced drivers (just using UK clips), there is no indication that experienced drivers are better at detecting hazards. Indeed, UK novice drivers with UK clips are slightly more accurate *M* = 87.2% , 95% CI [78.2%, 93.7%], than experienced drivers, *M* = 85.0%, 95% CI [74.9%, 92.3%], though this difference was not statistically significant, G^2 (1) = 0.3, *p* = .56. While this lack of significance contradicts many studies that have previously demonstrated such HP tests to discriminate between experienced and inexperienced drivers (e.g. Wallis & Horswill, 2007; Horswill et al., 2008; Deery, 2000), we have already noted in the introduction that failure to find this effect is not without precedent (e.g. Sagberg, and Bjørnskau, 2006; Lim et al., 2013; Yeung and Wong, 2015). The previous studies that were successful may also be over-reliant on statistical analyses that ignore variability in stimuli (and hence have inflated Type I error rates).

In addition to the failure to find experiential differences, several other interesting findings were noted that suggest the typical HP approach might be culturally sensitive. First, it was notable that Chinese drivers made fewer hazard responses over all the clips and especially for the UK clips, compared to the other two groups. In contrast, both Spanish and UK participants produced a greater number of extra hazard responses. This suggests that Chinese participants seem to be less sensitive to (or more accepting of) hazards from all three countries. Chinese drivers were also slower to respond to hazards across all three countries compared to the other drivers, while the UK drivers were the fastest. The slow responses of Chinese drivers may stem from their high threshold for reporting hazards, which seems apparent in their lower frequency of hazard responses. Chinese drivers are continuously exposed to a higher frequency of potential hazards which supports the hypothesis that criterion bias may influence the simple push-button response required in the traditional hazard perception methodology. Conversely, the faster responses of the UK drivers may reflect their previous exposure to the national UK test.

Interestingly, we found a significant interaction between nationality and clip origin which showed that Chinese drivers' extra hazard responses were particularly low for the Chinese clips. This suggests that Chinese drivers may indeed be more desensitized to hazards as they were not pressing as much as the other participants regardless of the more hazardous Chinese driving environment (Lim et al., 2013).

Chinese and Spanish clips evoked the greatest number of extra hazard responses per minute across all participants, suggesting that both Chinese and Spanish clips are more complex, and contain more precursors than the UK clips. This in itself is unsurprising as China has the highest collision rate of the three countries, and is therefore likely to have more potential hazards. Of greater interest is that there is no difference between the Chinese and Spanish clips which might suggest that both environments look equally hazardous to our paticipants, even though Spain typically reports fewer traffic accidents than China. In fact, there was a significant interaction between participants' nationality and clip origin where both Chinese and UK drivers rated Spanish clips as the most hazardous, while Spanish participants gave the highest ratings for the Chinese clips. This supports the notion that the UK driving environment is the one with the lowest level of on-road complexity, while the Spanish environment appears on a par with the Chinese one to our participants.

The high threshold bias of the Chinese drivers is also reflected in the hazardousness ratings provided by all participants following each clip. They gave the lowest ratings, followed by the UK drivers, with Spanish drivers providing the highest hazardousness ratings.

To summarise, this current hazard perception test does not appear appropriate to export to other countries. First, it does not differentiate between experienced and inexperienced drivers, which is considered to be a mainstay of test validity in the literature. While disappointing, this is not necessarily insurmountable. Some hazards are likely to be more successful in discriminating between experienced and inexperienced drivers (Crundall et al., 2012; Crundall, 2016) and it could be possible to collect new hazards that add to the validity of the test.

A second barrier to exporting a hazard perception test is that, in its current form, our test appears highly sensitive to cultural differences between our driver groups, which is considered to be a problem for test fairness (Allen and Walsh, 2000; Gesinger, 1992; Padilla and Medina, 1996). All three nationality groups were found to differ on various measures, suggesting that the traditional methodology cannot simply be transplanted to another country where driving norms, social rules, and on-road complexity may all differ.

Finally, the traditional test is potentially confounded by a number of issues that have been raised in this study, including criterion bias, or the individual threshold of drivers for judging something to be hazardous. Individual thresholds can be influenced by cultural differences in acceptable driving norms (e.g. Lim et al., 2013, 2014), and by driving experience and expertise, with more advanced drivers discounting hazards if they fall within their self-perceived range of skill (e.g. Crundall et al., 2003). A second potentially confounding issue can be seen in the correlation between accuracy in responding to target hazards, and the overall number of extra hazard responses per minute. Though relatively small, this relationship suggests that responding more frequently is linked to greater accuracy in identifying hazards. While this may also be linked to experience (as experienced drivers make more EHR/m to Chinese clips than inexperienced drivers), more frequent clicking may result in some responses falling within the scoring window by chance rather than reflecting identification of the *a priori* hazard.

This raises further issues of how one defines scoring windows. There are no accepted guidelines on what should constitute a hazard onset or offset. Relatively tight scoring windows are required, if one simply relies on a non-locational hazard click, in order to minimise the probability of misattributed responses occurring in the hazard window. Unfortunately, reducing the scoring window length to limit false alarms, increases the probability of missing correct responses (e.g. early hazard responses from highly experienced drivers).

An alternative solution to the scoring window problem is to include a measure of accuracy. For instance, instead of simply pressing a button when one sees a hazard, the participant might have to indicate where that hazard occurred via a touch-screen press or a locational mouse click (e.g. Banbury, 2004; Wetton et al., 2010, 2011). Both of these methods have potential drawbacks however such as individual differences in pointing tasks (e.g. Zhai, Kong and Ren, 2004), age and gender differences in mouse and touch screen use (e.g. Hertzum and Hornbaek, 2010; Wahlström et al., 2000; Yamauchi et al, 2015), and possible systematic differences between experience groups that may affect the speed-accuracy relationship. For example, if experienced drivers spot hazards earlier than inexperienced drivers (e.g. Crundall et al. 2012), then the hazard is likely to be smaller (i.e. further away) than when spotted by inexperienced drivers. According to Fitts’ Law (Fitts, 1954), a smaller target will increase demands on accuracy and therefore slow pointing speed, potentially negating the experiential benefit of perceiving the hazard sooner. Nonetheless, we cannot dismiss the research that has shown significant experiential differences using this response mode, and it remains an exciting option worth pursuing.

In conclusion, our current hazard perception test, based on the traditional UK methodology, produces more differences between groups on the basis of nationality than driving experience, and is influenced by the context of clips. We recognise that this is not the only HP methodology that we could have implemented, and that the variations employed by many other researchers may have produced a better test. However, when considered alongside the problems of criterion bias, and issues related to the measurement and interpretation of simple response times, there appears to be little evidence that allows us to commend the export of this particular hazard perception test methodology to other countries. Rather than creating a culturally agnostic test of drivers' higher-order cognitive skills, we have created a culturally sensitive measure that cannot yet differentiate between safe and less-safe drivers based on experience.

Instead of the current flawed methodology, we need a new test that will tap into the expertise of drivers at spotting hazards that is independent of cultural background. At the same time, we need to remove both the problem of criterion bias, and the ambiguities of setting hazard-scoring windows. Finally, a new test should also address the lack of an accuracy measure by means that do not threaten systematically to mask any experiential benefit. To this end we have turned to a purer test of hazard prediction for the second experiment.

Experiment 2

The act of hazard perception contains a number of sub-processes including searching for hazardous precursors, predicting which hazard is most likely to occur, monitoring the prioritised locations, spotting and processing the eventual hazard, and then responding in a timely and appropriate manner. Indeed, the whole process of avoiding a hazard on the road is poorly reflected within the term ‘hazard perception’ and recently Pradhan and Crundall (2017) have argued that ‘hazard avoidance’ is a more appropriate overall term. While ‘hazard perception’ is not a broad enough term to capture the whole hazard avoidance process (such as selection of the most appropriate behavioural response; see Ventsislavova et al., 2016), neither does it confine itself to a perceptual process. We have noted evidence from experiment 1, and from other studies, which suggests that post-perceptual processes, such as comparison of the demands of the unfolding hazard to one’s own perceived skill, may influence the response. With such a nebulous definition of hazard perception, it is unsurprising to find that we are not completely clear on what the traditional HP test is measuring.

How do we overcome this problem of measuring hazard perception, or hazard avoidance, skill? There are two obvious alternatives. First, we might consider analysing the whole hazard avoidance process rather than just recording timed responses to hazards contained in video clips. This could be done naturalistically by fitting vehicles with cameras and sensors to monitor real-world driving behaviour (e.g. Dingus et al., 2006; Barnard et al., 2016), or by studying driver behaviour in a simulator (Chan, Pradhan, Pollasek, Knodler and Fisher, 2010; Crundall et al., 2010, 2012). While both methodologies have contributed significantly to our understanding of why drivers crash, they do not provide detailed understanding of the sub-processes involved, and they do not provide a suitable tool for mass testing.

A second alternative to overcome the problems inherent in the traditional HP methodology is to pinpoint a more specific sub-process that can be more precisely measured. Pradhan and Crundall (2017) have defined these different sub-process, one of which is the act of *hazard prediction*. This process is akin to Endsley's (1988a, 1995) third level of situation awareness: projection of future states and locations of objects on the basis of their current configuration and trajectories. The driver collects evidence from all potential hazard precursors and predicts whether any of them will come into conflict with her own vehicle. Should this process identify an imminent hazard, the driver prepares to act accordingly. We believe this sub-process lies at the heart of all hazard avoidance, and is likely to be the key skill that traditional hazard perception tests are imperfectly measuring. In order to assess this prediction skill more directly, the traditional hazard perception test can be simplified following the methods employed by the Situation Awareness Global Assessment Test (SAGAT; Endsley, 1988b). Rather than letting the clips play all the way through, clips in the *hazard prediction test* are cut short, occluding as soon as the hazard begins to develop. Instead of asking participants to make a timed response to the hazard, they are simply asked 'What happens next?', with their responses coded as correct or incorrect. This rests on the assumption that safer drivers know where to look for precursors to potential hazards, and can process, prioritise and monitor these precursors accordingly, giving them the best possible chance of looking in the right place at the right time (i.e. looking at the precursor just as it begins to develop into a hazard before the screen is immediately occluded). Less-safe drivers are less likely to be looking in the most appropriate locations and will therefore have a reduced chance of predicting the hazard.

This purer measure of hazard prediction skill offers several advantages over the traditional hazard perception methodology. First, it provides a measure of accuracy that is unavailable to traditional hazard perception tests (without some form of hazard localisation in the response, which may bring with it a new set of confounds). Secondly, it removes the need for temporal scoring windows which may penalise very good drivers who press slightly too soon. Thirdly, it removes the controversy of dealing with missing response time data. The traditional approach of recording the maximum possible RT in otherwise empty cells (McKenna et al., 2006) has been argued to distort results (cf. Parmet, Meir and Borowsky, 2014, who recommend the use of survival analysis). We addressed this by treating the data as right-censored in Experiment 1. Such an approach incorporates additional uncertainty for the censored data and thus may require larger data sets to detect effects as well as being more complex. The hazard prediction test avoids this problem by dropping RTs as the main measure.

A fourth benefit is that it removes the possibility that the test instructions are interpreted differently across the cultures. We know that terms like “hazard” and “hazardousness” are inherently prone to individual differences in interpretation (Wetton et al., 2011), and thus cultural differences are highly probable. Despite our best efforts in the first study (forward-backward translation, having Chinese and Spanish researchers run the experiments in their respective countries), our participants may have had significantly different understanding of what constitutes a hazard. With the hazard prediction test however, we remove this problem by simply asking “What happens next?”.

Finally, the hazard prediction test should remove criterion bias. There is no implicit or explicit motivation for participants to compare an unfolding hazard to their own self-perceived skill when responding. Instead, they simply report what happens next, regardless of how hazardous they believe the imminent event would be for them personally (though self-perceived hazardousness can still be captured after they have made the prediction) . If the cultural sensitivity of the test used in Experiment 1 is, at least in part, due to the confounding of criterion bias with the traditional timed hazard response, then a new test based just on this prediction element of the skill may be more robust (Jackson et al., 2009; Castro et al 2014; Lim et al., 2014; Crundall, 2016).

The hazard prediction test for experiment 2 was created using the same clips employed in experiment 1. The clips were edited to cut to a black screen as soon as the hazard begins to appear. Following occlusion, participants typed their responses to what they believed would happen next. A new cohort of experienced and inexperienced drivers was recruited across the three countries for this second experiment. We predicted that the prediction test would be more successful than the hazard perception test in discriminating between the driver groups, and that the test would demonstrate fewer cultural sensitivities.

Method

Participants

A hundred and fifty-three participants took part in Experiment 2. The sample was composed of 50 Chinese, 52 Spanish and 51 UK drivers. One participant was later excluded (from the UK sample) due to difficulties categorising the individual as experienced or inexperienced. All of the participants held a full or a learner-driver licence from their country. Participants were split into two sub-groups of experienced and inexperienced drivers following the method used in Experiment 1. In China we recruited 26 experienced drivers (mean age of 25.3, an average of 5 years of post-licensure experience, and a mean annual mileage of 5474 miles) and 24 inexperienced drivers (mean age of 22.7, an average of 1 year of post-licensure experience, and a mean mileage of 33.7 miles). In Spain we recruited 27 experienced drivers (mean age of 40.9, an average of 21 years of post-licensure experience, and a mean mileage of 20183 miles) and 25 inexperienced Spanish drivers (mean age of 20.2, an average of 1 year of post-license mean experience and mean mileage of 28.9 miles). In the UK 23 experienced UK drivers were recruited (mean age of 24.4, an average of 7 years of post-license mean experience and mean annual mileage of 5587 miles), along with 27 inexperienced UK drivers (mean age of 19.4, an average of 1 year of post-license experience, and a mean annual mileage of 266.7). Across all countries, the mean age of experienced drivers was 30.2 years, with an average of 11 years of post-licensure experience, and they had driven an average of 10415 miles in the previous year, while inexperienced drivers had a mean age of 20.8, with an average of 1 year of post-licensure experience, and had driven an average of 109.6 miles.

Participants from the three countries were recruited either from the respective Universities or from local driving schools. All of the participants were volunteers.

Materials and apparatus

The apparatus and stimuli for this experiment were the same as those used in experiment 1, though the video clips were edited to stop immediately prior to the appearance of the hazard for the current experiment (immediately following hazard onset), with the clip occluded by a black screen. The edited clip always gave enough information for participants to deduce what would happen next in the driving scene providing they were looking in the appropriate location just before occlusion (Jackson et al., 2009). At the end of each clip a black screen was displayed. The duration of the clips varied between 12 ms and 58 ms.

As an example, consider clip 1 from the Chinese block (see Table 1). In this clip (as used in experiment 1) a pedestrian looks to cross the road from the right but is then obscured by a turning vehicle. When the vehicle has finished the manoeuvre, the pedestrian is already crossing the road in front of you. For the current hazard prediction test, this clip was edited to end in the middle of the obscuring vehicle's manoeuvre, at a point where part of the hazardous pedestrian emerging in the road can be seen. An experienced driver should notice the pedestrian before the vehicle turns, and therefore should monitor the trailing edge of the obscuring vehicle to assess whether the pedestrian has indeed entered the road. The briefest glimpse of the re-emerged pedestrian is only likely to be spotted if the driver is aware of the unfolding hazard and is actively seeking the pedestrian.

Design and Procedure

The design of the study was identical to that of experiment 1, except for the dependant variable. Instead of a response time measure to the hazard, the screen occluded immediately prior to the hazard fully developing, and participants were asked to type what they thought happened next into a text entry box on the screen.

Upon entry to the lab all participants were first required to fill in the demographic questionnaire and were then seated 60 cm from the screen and viewed the instructions in their native language. They were told that they were going to see 30 video clips from three different countries. They were asked to watch each clip carefully because at some point the clip would end and be occluded by a black screen. They were further instructed that, following occlusion, an on-screen question would ask them ‘What happens next?’ At this point they were told they should type a short answer, describing how the driving situation was going to develop. Participants were informed that the entry box was limited to 150 characters and were therefore encouraged to keep their responses brief and to the point. Participants typed their answers in their native language which were later translated into English for coding. To focus their responses, participants were encouraged to report any source of potential hazard, its location on the screen at the point of occlusion, and how the situation was about to develop (e.g. ‘A pedestrian behind the turning car on the right is about to step into the road’). Before the start of the actual experiment, participants viewed a practice trial, where they had the opportunity to familiarize with the experiment and ask any questions. They were given feedback on their answer in the practice trial (by viewing the full clip once they had provided a response), but not in the main study. When participants were comfortable with what they were required to do, they began the experiment.

Typed responses were later coded with one point given for each correct answer (ideally specifying what and where the object of interest was, and how the event would unfold) and zero points for an incorrect answer. Where participants failed to report the three suggested items in their answer, but it was still unambiguously correct, they were still awarded the point. For example, if a clip stopped at a point where a pedestrian was approaching a zebra-crossing and looked at the film car, an ideal correct answer would be “A pedestrian from the left is about to cross the road” (table 1). However if there were no other pedestrians in the scene, an answer that omitted to note that the pedestrian was on the left, would still receive a point.

Once they had provided an answer, participants were presented with an on-screen Likert scale, ranging from 1 to 7, to report how hazardous they felt the clip was (with ‘not at all hazardous’ to ‘extremely hazardous’ as the anchors). The number on the scale was selected via a mouse click. Following this response, a one second fixation cross was presented before the next clip started. At the end of the block, there was a brief pause before the next block would begin. The order of the blocks was randomised (i.e. which country’s clips they saw first) and the order of the clips within the block was randomised.

Results

For this analysis 152 participants were included. One participant from the UK was removed due to difficulty in classifying her as either experienced or inexperienced (having obtained driving licence in 1998, but reporting extremely low mileage).

To test whether there were differences in the accuracy of hazard prediction performance across the factors, the 2x2x3 factorial design was analysed as in Experiment 1 using a multilevel logistic regression with participant and clip as random factors. The between-groups factors were the experience level of drivers (experience vs. inexperienced drivers) and their nationality (Chinese vs. Spanish vs. UK). The within group factor was the origin of the clip (China vs. Spain vs. UK). An intercept only model (with no predictors) estimated the *SD* of the participant random effect as 0.804 and the *SD* of the clip random effect as 1.278 indicating that only 28% variation at level 2 of the model is attributable to participants – with variability in clips accounting for the majority (72%) of level 2 variance. This indicates that a traditional ANOVA analysis – that treats variation between clips as zero – would substantially underestimate standard errors. The deviance (likelihood ratio Chi Square, G^2) for the intercept only model was 5023.8 and decreased to 5012.9 for a model including main effects of nationality, experience and clip type. This improvement in model fit was not statistically significant, ΔG^2 (5) = 9.9, p = .077. However, G^2 decreased dramatically for a model with all two-way interactions, ΔG^2 (8) = 24.2, p < .002, with a negligible improvement with the addition of the three-way interaction, ΔG^2 (4) = 2.8, p > .05. The two-way interaction model therefore appears to be the most informative.

The pattern of accuracy across all conditions is shown in Figure 7. A main effect of drivers' experience was found, ΔG^2 (1) = 7.1, p < .01, *OR* = 1.48, 95% CI [1.10, 1.99]. Experienced drivers, *M* = 50%, 95% CI [38%, 63%], were on average more likely to predict the hazards than novices, *M* = 41%, 95% CI [29%, 53%]. No main effects were detected for clip origin, ΔG^2 (2) = 1.2, p = .56, or participant nationality, ΔG^2 (2) = 2.2, p = .33.

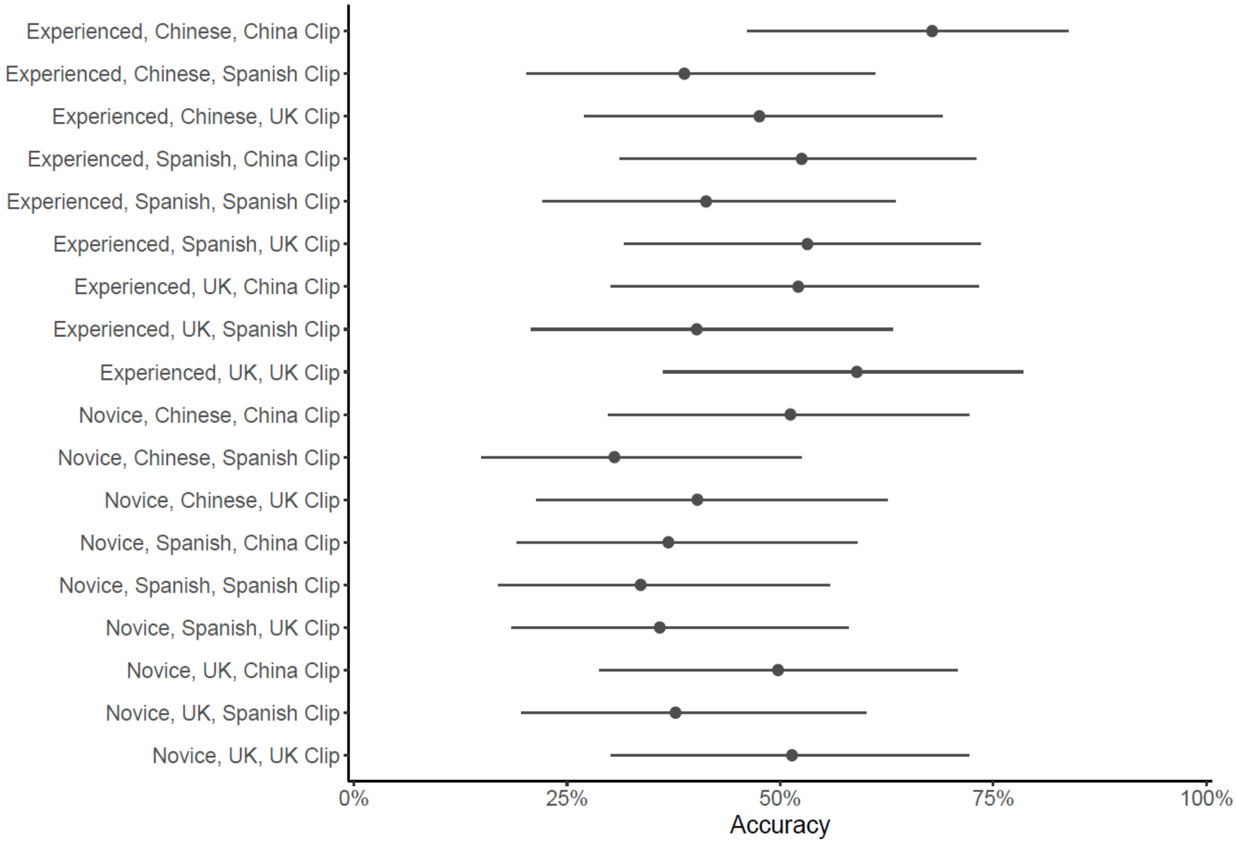


Figure 7. Percenatges of accuracy in prediction across driving groups, participant's nationality and clips (with error bars).

A nationality by clip origin interaction was also detected, ΔG^2 (4) = 21.4, p < .001 (see Figure 8). The pattern of accuracy across these conditions is complex – but generally performance is superior for drivers when the clip origin is consistent with the participant nationality (with the exception that UK drivers are slightly better with the Spanish clips than Spanish drivers). To confirm this we followed up the significant interaction with an interaction contrast comparing own nationality clips with other nationality clip conditions. This contrast was statistically significant, G^2 (1) = 16.7, p < .001, likely explaining the bulk of the variation in accuracy contributing to the interaction.

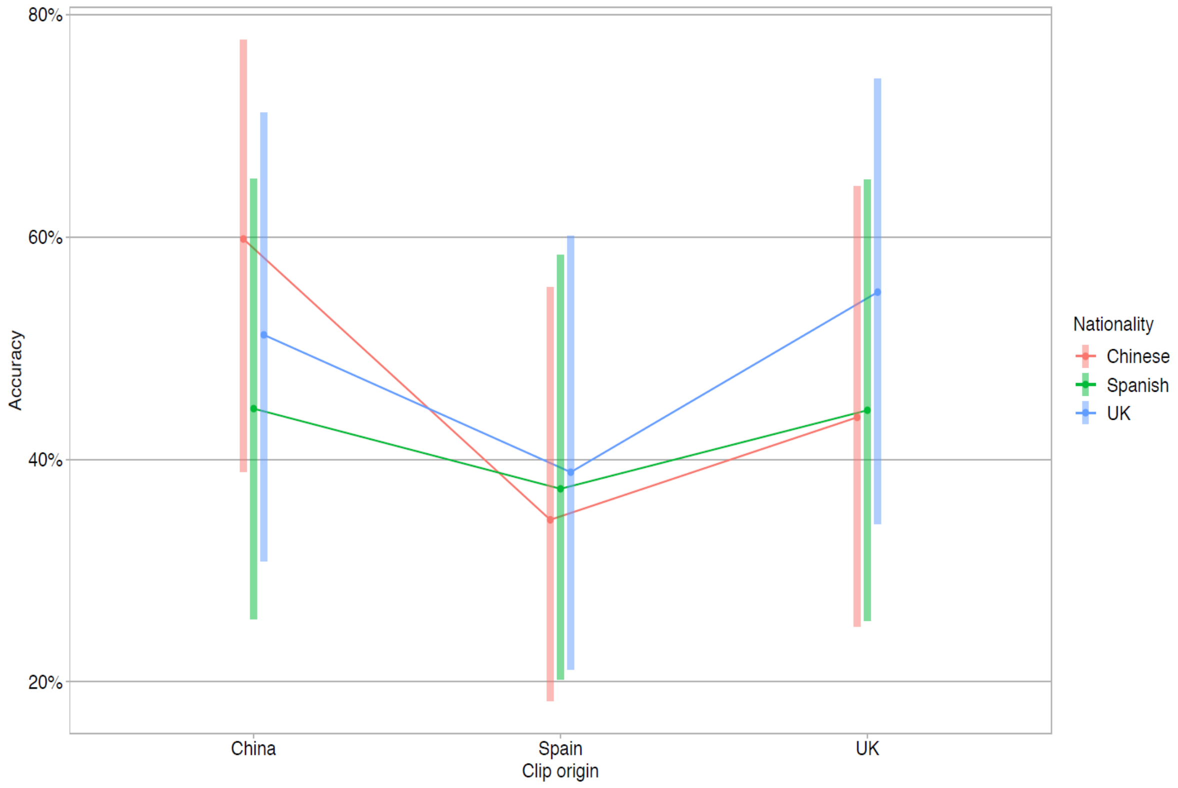


Figure 8. Accuracy in prediction across different nationality vs clip origin (with error bars).

Hazardousness ratings for hazard prediction

As with the hazard perception study, participants were asked to provide a hazardousness rating on a scale of 1 to 7 (where higher numbers reflect greater levels of perceived hazardousness). Again a multilevel ordinal logistic regression was used to analyse the ratings using a factorial design (with experience, nationality and clip origin as factors). The intercept only model (G^2 (8) = 15252) was a worse fit than a model with all main effects (ΔG^2 (5) = 29.3, $p < .0001$), which in turn was a worse fit than a model with all two-way interactions (ΔG^2 (8) = 35.4, $p < .0001$). Adding the three-way interaction only marginally further improved the model (ΔG^2 (4) = 5.8, $p = .21$).

There was a main effect of clip origin, G^2 (2) = 8.8, $p = .012$. Chinese clips were rated as more hazardous than Spanish Clips (4.26 vs. 3.73) with a smaller difference between the

Spanish clips and the UK clips (3.73 vs. 3.63). Hochberg corrected post hoc tests revealed significant differences between the Chinese and Spanish (adjusted *p* = .025) and UK clips (adjusted *p* = .008) but not between Spanish and UK (adjusted *p* =. 61) . Nationality also produced a significant main effect, G^2 (2) = 20.2, *p* < .0001, with Spanish drivers giving the highest ratings (4.24) followed by UK drivers (4.02), with Chinese drivers giving the lowest hazard ratings (3.33). Hochberg corrected pairwise tests revealed Chinese drivers to give lower ratings than both Spanish and UK drivers (adjusted *p* < .0001 and *p* < .002 respectively), with no difference between Spanish and UK drivers (*p* = .26).

In regard to driver experience, though experienced drivers were not found to give significantly different ratings to those provided by novices overall (3.86 vs. 3.88; G^2 (1) = 0.2, *p* = .64), driver experience did interact with clip origin (G^2 (2) = 12.8, *p* = .002). Both driver groups rate Chinese clips as more dangerous than Spanish clips, but this effect is more pronounced in the experienced driver group. There was also an interaction between clip origin and nationality, G^2 (5) = 15.1, *p* < .005. This largely followed the pattern obtained for the main effects except that the UK participants rated Spanish clips more hazardous than UK clips (adjusted *p* < .05).

Discussion

Unlike the hazard perception test of experiment 1, the hazard prediction test successfully discriminated between experienced and novice drivers, with the experienced drivers outperforming the inexperienced across all nationalities. These results are consistent with the limited previous research, demonstrating that the prediction test is a more robust discriminator of driver experience than the traditional hazard perception test (Lim, 2014, Castro, 2014, Crundall, 2016). The superiority of the hazard prediction test is all the more convincing in that it discriminated between our driver groups using the same clips as the unsuccessful hazard perception test in experiment 1. In addition, we did not find any interaction between experience and participant nationality demonstrating that the prediction test is less sensitive to cultural differences than the hazard perception test.

There was however a significant interaction between clip origin and nationality, showing that performance was better when the clip origin is consistent with participants's nationality. This suggests that the familiarity of potential precursors available in the environment might influence the ability to identify the correct target. Being more aware of what possible hazards one might find aids the detection of early precurosrs (Crundall, 2016; Underwood. Chapman, Bowden and Crundall, 2002). Although this might suggest that hazard prediction is affected by context, it did not influence prediction accuracy between the experienced groups. It should be noted, that UK drivers were actually slightly better for the Spanish clips than Spanish drivers meaning that the the type of hazard (regardless of context) may influence performance, too (although we did not find a main effect for clip origin). Thus, the finding that some drivers perform better when viewing clips filmed in their own country, does not detract from the claim that the prediction test is a more culturally agnostic form of assessment than the hazard perception test.

There were however still differences between the hazardous ratings in regard to clip origin in contrast with Experiment 1. Chinese clips were rated as most hazardous, followed by the Spanish and UK clips. However, participants in this study did not see the materialised hazards which means that the Likert scores could be reflecting general visual clutter, complexity and congestion, rather than the *a priori* hazard in particular. Participants may have presumably referenced other potential hazards that they had seen in the clip in order to provide a hazard rating. As UK clips evoked the least extra hazard responses in experiment 1, it is safe to conclude that these clips contain less potential hazard precursors, and this fact has also been reflected in the ratings in experiment 2.

Comparison of the two tests

It is possible to directly compare the performance on the two tests, using the prediction accuracy from experiment 2 and the percentage of hazards that received a timed button response in experiment 1 (though note that we cannot claim that all responses that fell in the scoring window in the hazard perception test were referencing the actual hazard – this is one of the problems with the traditional HP methodology). In the analysis reported below we only focus on the main effect of *test type* (whether accuracy scores differ across the hazard *perception* and hazard *prediction* tests), and any emerging interactions with test type. Accuracy rates for the two tests were compared with a 2x2x3x3 factorial design using a multilevel logistic regression model with the factors of test type (Perception vs. Prediction), participant experience (inexperienced vs. experienced), participant nationality (Chinese, Spanish or from the UK), and clip origin (China, Spain, UK). For these models the main effects only model was more informative than the intercept only model, ΔG^2 (6) = 198.1, $p <$ .0001, with the two-way interaction model improving model fit further still, ΔG^2 (13) = 76.7, $p < .0001$. The three-way interaction model also offered additional improvement in fit, but was not significantly better than the two-way model, ΔG^2 (12) = 20.8, $p > .05$, with almost no change in fit after adding the four-way interaction, ΔG^2 (4) = 1.4, $p > .05$. As the focus here is on differences between the tests we report only tests of effects comparing *hazard perception* to *hazard prediction.*

The results showed that there was a main effect for the type of test, G^2 (1) = 168.0, $p <$ .0001. Participants scored higher on average for the hazard perception test compared to the hazard prediction test (77.2% vs. 47.8%). A significant interaction was also found for test-type and experience, G^2 (1) = 168.0, $p < .0001$. Despite the prediction test appearing more difficult than the perception test, it is clear that the benefit of experience only holds for hazard prediction rather than hazard perception (see Figure 9).

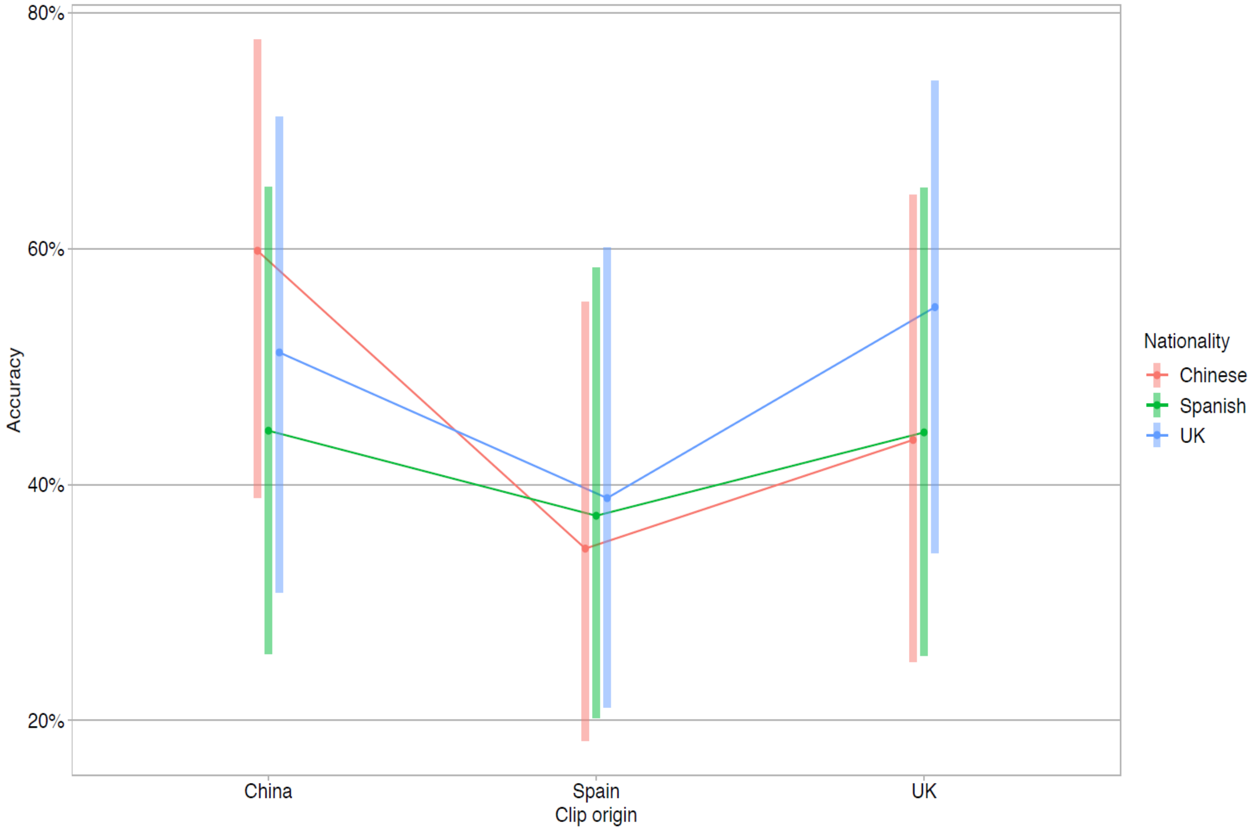


Figure 9. Percentages of prediction accuracy for test type across experience (with error bars).

A significant interaction was found for test type and nationality, G^2 (1) = 10.3, $p <$ .01. As can be seen from Figure 10, the variation in performance across the nationalities was significantly greater in the hazard perception test (reflected in the main effect of nationality found in Experiment 1), than in the hazard prediction test (with no significant main effort of nationality in Experiment 2).

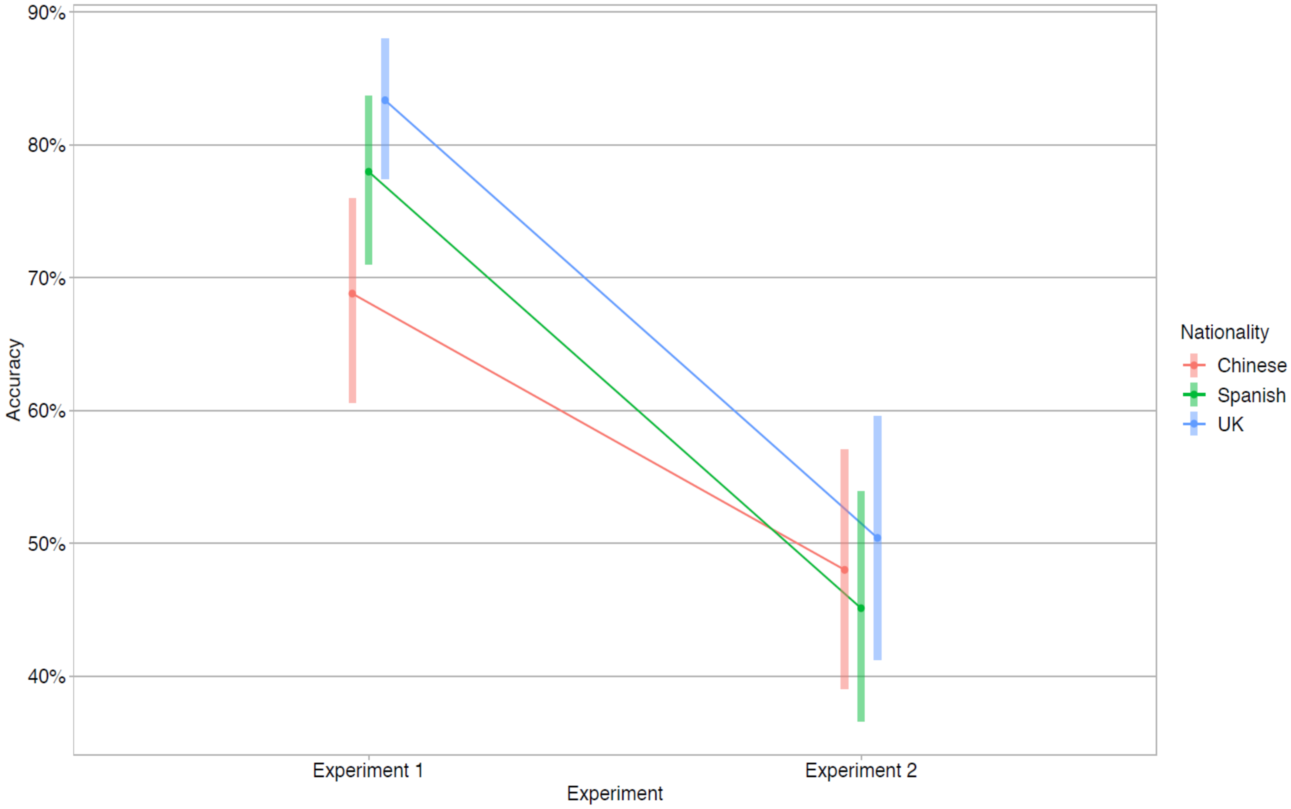


Figure 10. Percentage of prediction accuracy for test type across nationality (with error bars).

Finally there was a significant interaction between clip origin and test type, G^2 (1) = 35.4, $p < .0001$. Again, this effect captures the difference in clip origin main effects for the separate experiments. In the hazard perception test the Chinese hazards were the hardest to detect (72.4%) and UK hazards were the easiest (80.4%), however for the prediction test Spanish hazards appeared to be the hardest to predict (42.7%) and Chinese the easiest (51.7%). No other effects incorporating test type were statistically significant (all $p > .05$).

General Discussion

The aim of this paper was to assess whether two variants of hazard perception test are suitable for export to different driving cultures, acknowledging the possibility that the typical hazard perception test methodology may be culturally sensitive, and therefore less suitable for adoption in other counties. This was a novel endeavour as, though many research groups around the world have investigated hazard perception in their own countries, they have done so with vastly differing methodologies making it difficult to compare the validity of the tests across different regions. Only one previous attempt has been made to assess hazard perception skills of drivers from different countries using the same clip set, but the results of that study were inconclusive (Lim et al., 2013; 2014).

The results of the traditional HP test format revealed considerable differences in driver groups from different countries. For instance, the Chinese participants were the slowest to react to the hazards and identified significantly fewer hazards, and made fewer responses overall, in comparison to the Spanish and UK drivers. Conversely, UK drivers showed faster responses and identified the most hazards. Despite finding these effects of participant nationality, we failed to find significant differences between experienced and inexperienced drivers in regards to their hazards responses. Both groups performed identically in the test and, therefore, we cannot conclude that the traditional methodology of the hazard perception test is transferable to other countries, as we were not able to establish test-validity in the UK sample in the first instance.

Not only could we not find experiential differences, the cultural differences of our driver groups appeared to significantly influence the way they approached the test. Chinese drivers rated the clips as less hazardous than the other driver groups, which may account for their slower response times. They were also the least accurate and they produced significantly lower rate of extra hazard responses compared to the Spanish and UK drivers. This is ostensibly due to differences in cultural hazard thresholds. On the basis of the higher traffic collision statistics in China, compared to the UK and Spain, it is safe to assume that Chinese drivers are likely to encounter many more hazards on the road in every day driving. This increased exposure to hazards presumably desensitises the Chinese drivers to the relative seriousness of some hazardous events, increasing their thresholds for reporting them. This is most likely to be the cause of the slower response times in the traditional hazard perception test used in experiment 1.

In addition, a correlation was identified between the number of *a priori* hazards that participants responded to within the scoring window and the overall number of extra hazard responses that participants made. This raises a clear concern for the traditional hazard perception methodology, as it appears that the high performance of individuals may be influenced by clicks falling within the scoring window that do not necessarily reflect the *a priori* hazard.

While the current hazard perception test raised interesting questions regarding differences in the driving environment and the individual hazard thresholds, the results also suggest that the traditional hazard perception methodology would not be suitable for use in different countries, where environmentally-evoked high criterion bias may render the test insensitive to the skills of the safest drivers in those environments.

As the traditional hazard perception test failed to find differences between the experienced groups, an alternative hazard prediction test was created for experiment 2 based on initial studies that we had already conducted in the UK and Spain (Castro et al., 2014; Crundall, 2016; Jackson et al., 2009). The hazard clips were edited to occlude just as the hazard begins to develop, and participants were asked ‘What happens next?’.

Crucially, the clips edited for the hazard prediction test were the same as those used in the hazard perception test, allowing a direct comparison of the two tests. This is the first time that the hazard perception and hazard prediction tests have been directly compared in a single analysis[4], though this was complicated by the fact that the two tests record very different primary measures: response times and percentage accuracy, respectively. However, as the

[4] We have since compared hazard perception and hazard prediction test variants using video clips filmed from fire appliances on blue-light training runs. Once again, we found the prediction test to be the better discriminator of driver groups (Crundall and Kroll, 2018).

hazard perception test required response times to fall within a temporal scoring window around the appearance of the hazard, the presence or absence of a response allowed the calculation of an accuracy score that could be compared to the hazard prediction test.

While participants found the hazard prediction test much harder than the hazard perception test, the superiority of the prediction test in discriminating between driver groups on the basis of experience was clearly demonstrated. Most importantly, the main effect of experience, with experienced drivers outperforming novices, was present across the participants as a whole, and did not interact with nationality. As the prediction test was designed to remove criterion bias, it was comforting to note that the cultural differences that arose in experiment 1, which were interpreted as potentially arising from hazard threshold differences, were ameliorated to a large extent in experiment 2.

Do different countries produce different hazards?

In both studies, we noted differences in extra hazard responses or ratings to the clips on the basis of their origin. This is unsurprising, as Beijing, Granada and Nottingham, differ on a great many characteristics. The higher population, congestion and collision rates in China suggest that this should provide the most hazardous stimuli. While the clips were filmed with the same protocol, there were inevitable differences in the visual clutter and frequency of hazardous precursors across the countries. From an experimental design point of view, this was not a great concern. As every participant saw clips from all three countries, we could thus analyse the relative differences between the responses of our participants across the three nationality groups.

The effect of clip origin on ratings in experiment 2 closely mirrored the behavioural findings in experiment 1 regarding the UK clips. Generally, the UK driving environment is considered to be the least dangerous. UK clips were rated as the least hazardous and received the lowest rate of extra hazard responses (although most of the time there were no significant differences between the Spanish and UK driving environment). The Chinese clips were rated as the most hazardous, presumably due to a greater number of precursors resulting in the possibility that participants considered the environment as too demanding and cluttered.

Both experiments yielded significant interactions between nationality and clip origin regarding accuracy. While in experiment 1 Chinese drivers were observed to perform particularly poorly on the UK clips, in experiment 2 we observed a familiraty effect. In the prediction test drivers performed better when clips were from their own country. It is understandable that participants are more accurate at identifying precursors in a familiar environment as they know where to look and what cues to serach for (Groeger, 2000). Despite this, the test still discriminated successfully between the experienced groups which indicates that familiarity is not influencing the overall purpose of the test. However, these results also indicate that it could be highly beneficial to expose drivers to hazards from different countries as they can be trained to identify precursors that maybe specific to particular environments. For drivers who cross national borders, and drive in a wide variety of cultural contexts, training in precursor identification in different countries may improve safety (for example, in long-haul HGV drivers).

Despite the differences in responses in experiment 1 across participant nationality and the familiarity effect in experiment 2, it was notable that many of the hazards across countries shared commonalities. Vehicles emerging from side roads, pedestrians crossing in front of the film car, and parked vehicles moving off, were all examples of *a priori* hazards that appeared across the countries. China did however produce many more overtaking hazards during filming than the UK (with 4 such hazards included in the final Chinese clip selection, and only two in the UK clip set). It is possible that we have previously underestimated the potential for overtaking hazards to be included in UK hazard perception tests, limited as we were by the self-imposed constraints of a single forward-facing perspective (i.e. without mirror information available to the participant). Thus, while the frequency with which hazards and precursors might occur ostensibly changes across countries, it is easy to identify *a priori* hazards that have a similar structure regardless of their origin.

This raises the possibility of developing a cohesive and culturally agnostic typology of hazards. Some attempts have been made in the literature to distinguish between coarse categories of hazards (e.g. latent vs. overt hazards; developing vs. abrupt hazards; behavioural vs. environmental prediction hazards; see Pradhan and Crundall, 2017, for a review), but there is an opportunity to classify hazards at a finer level. It is likely that some hazards will be more effective discriminators of driver safety than others (e.g. Crundall, et al., 2012; Crundall, 2016). If a hazard typology can have a degree of consistency across cultures, then this increases the value of developing such a system of categorisation.

The limitations of hazard prediction tests

The current studies suggest that the hazard prediction test is a better discriminator of driver safety than the hazard perception test, it is not without its limitations. For instance, it may be argued that the average experienced-driver score of 47.8% accuracy is not very high. We counter, however, that it is the difference between the two groups that is more important, rather than an absolute score. While the difference between experienced and inexperienced drivers was significant, this could be improved with iteration of the stimuli sets, as would occur in the development of a formal test.

Critics may also argue that, while we have criticised the hazard perception test for its reliance on ill-defined hazard onsets, the hazard prediction test similarly relies on an *a priori* hazard onset for deciding upon occlusion points. While this is true, the precise timing of the occlusion point appears to have little effect on the validity of the test, at least within certain parameters (Crundall, 2016; Ventsislavova and Crundall, 2018).

One other limitation is that the hazard prediction test only reflects one sub-component of a behavioural chain that allows a driver to spot, assess and safely respond to a hazard on the road (Pradhan and Crundall, 2017). We are aware that this pure measure of hazard prediction does not necessarily reflect the full ability of a driver to successfully avoid a hazard. There may be drivers who may have excellent abilities to predict, and therefore spot, hazards on the road, but whose threshold for responding to hazards is so high, that they are still considered to be at high risk of a collision. These drivers may simply be culturally desensitised to hazards. Alternatively, some individuals may have a high threshold for responding due to high-regard for their own skills, perhaps mixed with a desire to ‘teach a lesson’ to other drivers who transgress safety boundaries (e.g. braking at the last moment to maximise the apparent danger caused by the other driver, to demonstrate how hazardous the other driver’s actions were). The hazard prediction test will not identify these problems (and is not designed to).

If drivers’ individual thresholds for reporting a hazard are considered important enough to warrant assessment (and we believe they are), they should be measured independently of the ability to predict the hazard. Currently, the traditional hazard perception methodology confounds hazard prediction and hazard processing with hazard appraisal (Pradhan and Crundall., 2017) and thus does not provide an ideal assessment of any of these sub-components. We recommend that each sub-component of the whole *hazard avoidance* process be assessed by individual measures, including a separate assessment of the choice and extent of the response (e.g. harsh braking, slight adjustment to lane position, etc.). This will allow better understanding of how drivers differ in their responses to hazards, at different stages of the hazard-avoidance behavioural chain, as set-out by Pradhan and Crundall (2017).

One final point to note is that the free-response format of the current hazard prediction test does not lend itself to widespread automated testing, due to the lack of immediate feedback, and the possibility of coding errors and subjectivity influencing the scoring. This is an issue that we have addressed in another paper, with the development of a multiple-choice question format that retains the ability to discriminate between driver groups, while increasing the potential for testing on a national scale (Ventsislavova and Crundall, 2018).

Conclusions

Many researchers agree that hazard perception skill is, perhaps, the only higher-order cognitive skill to relate to crash-risk, and that hazard perception tests have huge potential for reducing collisions around the world. Despite this, researchers often disagree on the underlying process of hazard perception, and how to measure it accurately, with different research groups each adopting slightly different methodologies. This has made it impossible to assess the cross-cultural generalisability of hazard perception tests.

These studies represent the first large-scale attempt to compare identical methodologies across three countries. The results have shown the typical hazard perception methodology to be unreliable and sensitive to cultural differences. The hazard prediction test, however, demonstrated a clear experiential difference, and appeared more impervious to the nationality of participants. The superiority of the hazard prediction test was all the more convincing in that it used the same clips as those presented in the hazard perception test.

The results provide a clear steer that the hazard perception process involves criterion differences that appear culturally-biased, and that such threshold effects confound the traditional response time measures of hazard perception tests. The hazard prediction test provides a purer, culturally agnostic variant of the traditional hazard perception test, and offers a blueprint for future test development at a global level.

References

Allen, J., & Walsh, J. A. (2000). A construct-based approach to equivalence: Methodologies for cross-cultural/ multicultural personality assessment research. In R. H. Dana (Ed.), Handbook of cross-cultural and multicultural personality assessment. Personality in clinical psychology series (pp. 63-85). Mahway, NJ: Lawrence Erlbaum Associates.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioural sciences*. London: Palgrave Macmillan.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67(*1), 1-48.

Barnard, Y., Utesch, F., van Nes, N., Eenink, R., and Baumann, M. (2016). The study design of UDRIVE: the naturalistic driving study across Europe for cars, trucks and scooters. *European Transport Research Review, 8*(14), 1-10.

Borowsky, A., Oron-Gilad, T., Meir, A., Parmet, Y. (2012). Drivers' perception of vulnerable road users: A hazard perception approach. *Accident Analysis and Prevention, 44*, 160–166.

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.

Bürkner. P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80(1),* 1-28. doi:10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software 76(1).* DOI 10.18637/jss.v076.i01

Castro, C., Padilla, J.L., Roca, J., Benítez, I., García-Fernández, P., Estévez, B.,López-Ramón, M.F., Crundall, D., (2014). Development and validation of the Spanish hazard perception test. *Traffic Injury and Prevention, 15*(8), 817–826. http://dx.doi.org/10.1080/15389588.2013.879125.

Castro, C., Ventsislavova, P., Peña-Suárez, E., Gugliotta, A., García-Fernández, P., Eisman, E., Crundall, D., (2016). Proactive listening to a training commentary improves hazard prediction. *Safety Science, 82*, 144–154. http://dx.doi.org/10.1016/j.ssci.2015.09.018.

Chan, E., Pradhan, A. K., Pollatsek, A., Knodler, M. A., and Fisher, D. L. (2010). Are driving simulators effective tools for evaluating novice drivers' hazard anticipation, speed management, and attention maintenance skills? *Transportation Research Part F, 13*(5), 343-353.

Chapman, P., Underwood, G., Roberts, K., (2002). Visual search patterns in trained and untrained novice drivers. *Transportation Research Part F, 5*, 157–167.

Cheng, A.S.K., Ng, T.C.K., Lee, H.C. (2011). A comparison of the hazard perception ability of accident-involved and accident-free motorcycle riders. *Accident Analysis and Prevention, 43*(4), 1464-1471.

Christensen, R. H. B. (2018). ordinal - Regression Models for Ordinal Data. R package version 2018.4-19.

Chun, M.M. (1997). Temporal binding errors are redistributed by the attentional blink. *Perception and Psychophysics, 59*(8), 1191-1199.

Cocron, P., Bachl, V., Früh, L., Koch, I., Krems, J.F. (2014). Hazard detection in noise-related incidents - The role of driving experience with battery electric vehicles. *Accident Analysis and Prevention, 73*, 380-391.

Crundall, D. (2016). Hazard prediction discriminates between novice and experienced drivers. *Accident Analysis and Prevention, 86*, 47-58. http://dx.doi.org/10.1016/j.aap.2015.10.006.

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

Crundall, D., Chapman, P., Phelps, N., Underwood, G., (2003). Eye movements and hazard perception in police pursuit and emergency response driving. *Journal of Experimental Psychology: Applied, 9*(3), 163–174. http://dx.doi.org/10.1037/1076-898X.9.3.163.

Crundall, D., Crundall, E., Clarke, D., Shahar, A. (2012). Why do car drivers fail to give way to motorcycles at t-junctions? *Accident Analysis and Prevention, 44*(1), 88-96.

Crundall, D., Chapman, P., Trawley, S., Collins, L., van Loon, E., Andrews, B.,Underwood, G., (2012). Some hazards are more attractive than others. *Accident Analysis and Prevention, 45*, 600–609.

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

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

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

Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J., Peres, M. A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z. R., Jermeland, J., Knipling, R. R. (2006). The 100-car naturalistic driving study, Phase II: results of the 100-car field experiment. *NHTSA report DOT HS 809 593.* NHTSA: Washington, US.

Dobkins, K. R., and Bosworth, R. G., (2001). Effects of set-size and selection spatial attention on motion processing. *Vision Research, 12,* 1501-1517.

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, http://www.dft.gov.uk/pgr/roadsafety/drs/novicedrivers/conference/theaustralianperspectiveonth4667 (retrieved 18.01.09).

Endsley, M. R., (1988a). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society 32nd Annual Meeting.* Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M.R., (1988b). Situation awareness global assessment technique (SAGAT). *Proceedings of the National Aerospace and Electronics Conference* (pp789–795). New York: IEEE.

Endsley, M.R., (1995). Towards a theory of situation awareness in dynamic systems. *Human Factors 37*, 32–64.

Fitts, P.M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*, 381-391.

Foss, R. D., Martell, C. A., Goodwin, A. H., & O'Brien, N. P. & Center U. H. S. R. (2011). Measuring changes in teenage driver crash characteristics during the early months of driving. Washington, DC:AAA Foundation for Traffic Safety.

Gao, L., Yu, X. T., and Hou, L. T. (2015). Research for the novice driver's cCapacity of hazard perception and response. In Y. H. Kim (Ed.), *Proceedings of the 2015 international conference on engineering management, engineering education and information technology, 36*, 355-360.

Geisinger, K. F. (1992). Fairness and selected psychometric issues in the psychological testing of Hispanics. In K. F. Geisinger (Ed.), Psychological testing of Hispanics (pp. 17-42). Washington, DC: American Psychological Association.

Groeger, J. A. (2000). Understanding driving: Applying cognitive psychology to a complex everyday task. New York, NY: Psychology Press.

Hertzum, M., and Hornbaek, K. (2010). How age affects point with mouse and touchpad: A comparison of young, adult and elderly users. *International Journal of Human-Computer Interaction, 26*(7)*,* 703-734.

Horswill, M. S., Anstey, K. J., Hatherley, C. G., and Wood, J. M., (2010). The crash involvement of odler drivers is associated with their hazard perception latencies. *Journal of the International Neuropsychological Society, 16, 5,* 939-944.

Horswill, M. S., Garth, M., Hill, A., Watson, M. O. (2017). The effect of performance feedback on drivers' hazard perception ability and self-ratings. Accident Analysis and Prevention, 101, 135-142.

Horswill, M.S., Hill, A., Wetton, M. (2015). Can a video-based hazard perception test used for driver licensing predict crash involvement? *Accident Analysis and Prevention, 82*(23), 213-219.

Horswill, M.S., Marrington, S., McCullough, C.M., Wood, J., Pachana, N.A.,McWilliam, J., (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., Taylor, K., Newnam, S., Wetton, M., Hill, A., (2013). Even highly experienced drivers benefit from a brief hazard perception training intervention. *Accident Analysis and Prevention, 52*, 100–110. http://dx.doi.org/10.1016/j.aap.2012.12.014.

International Test Commission (ITC) (2010). International test commission guidelines for translating and adapting tests. Download from: http://www.intestcom.org> [retrieved, July 26, 2012]

Isler, R.B., Starkey, N.J., Williamson, A.R., (2008). Video-based road commentary training improves hazard perception of young drivers in a dual task. *Accident Analysis and Prevention, 41*, 445–452.

Jackson, L., Chapman, P., Crundall, D. (2009). What happens next? Predicting other road users' behaviour as a function of driving experience and processing time. *Ergonomics, 52*(2), 154–164.

Judd C.M., Westfall J., Kenny D.A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem, *Journal of Personality and Social Psychology, 103*(1), 54-69

Lagroix, H.E.P., Spalek, T.M., Wyble, B., Jannati, A., Di Lollo, V. (2012). The root cause of the attentional blink: First-target processing or disruption of input control? *Attention, Perception, and Psychophysics, 74*(8), 1606-1622.

Lim, P.C., Sheppard, E., Crundall, D. (2013). Cross-cultural effects on drivers' hazard perception. *Transportation Research, Part F, 21*, 194-206.

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.

Malone, S. & Brünken, R. (2016). The role of ecological validity in hazard perception assessment. *Transportation research Part F: Traffic Psychology and Behaviour*, *40*, 91-103.

McCartt A.T., Shabanova V.I., & Leaf W.A. (2003). Driving experience, crashes and traffic citations of teenage beginning drivers. *Accident Analysis and Prevention, 35*(3), 311–320.

McGowan, A.M. & Banbury, S. (2004). Evaluating interruption-based techniques using embedded measures of driver anticipation. In: Banbury, S., Tremblay, S. (Eds.), A Cognitive Approach to Situation Awareness: Theory and Application. Ashgate, Aldershot, UK.

McDonald, C. C., Goodwin, A. H., Pradhan, A. K., Romoser, M. R. E., Williams, A. F., (2015). A Review of Hazard Anticipation Training Programs for Young Drivers. *Journal of Adolescent Health, 57, 1* (Supplement), S15-23.

McKenna, F.P., Crick, J.L., (1991). Ha*zard Perception in Drivers: A methodology for Testing and Training. Final Report*. Crowthorne, UK: Transport Research Laboratory.

OECD/ITF (2015). *Road Safety Annual Report 2015*. OECD Publishing, Paris. http://dx.doi.org/10.1787/irtad-2015-en (last accessed on 19/12/16).

Padilla, A. M., & Medina, A. (1996). Cross-cultural sensitivity in assessment: Using tests in culturally appropriate ways. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (Eds.), Handbook of multicultural assessment: Clinical, psychological, and educational applications (pp. 3-28). San Francisco: Jossey-Bass Publishers.

Palmer, J., Ames, C. T, and Lindsey, D. T., (1993). Measuring the effect of attention on simple visual-search. *Journal of Experimental Psychology: Human Perception and Performance, 19, 1*, 108-130.

Parmet, Y., Meir, A., and 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 y driving record of undergraduate males. Accident Analysis and Prevention, 6, 45–58.

Pradhan, A. K. and 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.

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). Retrieved from http://www.pstnet.com.

R Core Team (2018). R: A language and environment for statistical computing (Version 3.2.5) [computer program]. R Foundation for Statistical Computing, Vienna, Austria.

Renge, K., (1998). Drivers' hazard and risk perception, confidence in safe driving, and choice of speed. International Association of Traffic and Safety Sciences Research, 22, 103–110.

Rosenbloom, T., Perlman, A., Pereg, A. (2011). Hazard perception of motorcyclists and car drivers. Accident Analysis and Prevention, 43(3), 601-604.

Sagberg, F., Bjørnskay, T. (2006). Hazard perception and driving experience among novice drivers. Accident Analysis and Prevention, 28(2), 407–414.

Scialfa, C.T., Pereverseff, R.S., Borkenhagen, D. (2014). Short-term reliability of a brief hazard perception test. Accident Analysis and Prevention, 73, 41-46.

Shaffer, J. (1986). Modified Sequentially Rejective Multiple Test Procedures. *Journal of the American Statistical Association, 81*(395), 826-831.

Shahar, A., Alberti, C.F., Clarke, D., Crundall, D. (2010). Hazard perception as a function of target location and the field of view. Accident Analysis and Prevention, 42(6), 1577-1584.

Shahar, A., Van Loon, E., Clarke, D., Crundall, D. (2012). Attending overtaking cars and motorcycles through the mirrors before changing lanes. Accident Analysis and Prevention 44(1), 104-110.

Shimazaki, K., Ito, T., Fujii, A., and Ishida, T., (2017). Improving drivers' eye fixation using accident scenes of the HazardTouch driver-training tool. *Transportation Research Part F, 51*, 81-87.

Thomas, F. D., Rilea, S. L., Blomberg, R. D., Peck, R. C., Korbelak, K. T. (2016). Evaluation of the safety benefits of the risk awareness and perception training program for novice teen drivers (Report No. DOT HS 812 235). Washington, DC: National Highway Traffic Safety Administration.

Underwood, G., Chapman, P., Bowden, K., & Crundall, D. (2002). Visual search while driving: Skill and awareness during inspection of the scene,

Transportation Research Part F: Traffic Psychology and Behaviour, *5*(2), 87–97

Underwood, G., Ngai, A., Underwood, J., (2013). Driving experience and situation awareness in HP. *Safety Science, 56*, 29–35.

Vehtari, A., Gabry, J., Yao, Y., & Gelman A. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.0.0.

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

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.

Vlakveld, W.P. (2011). Hazard anticipation of young novice drivers. Proefschrift Rijksuniversiteit Groningen. SWOV-Dissertatiereeks, Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV, Leidschendam.

Vlakveld, W. P. (2014). A comparative study of two desktop hazard perception tasks suitable for mass testing in which scores are not based on response latencies. *Transportation Research Part F, 22,* 218–231. http://dx.doi.org/10.1016/j.trf.2013.12.013.

Wahlström, J., Svensson, J., Hagberg, M., and Johnson, P. W. (2000). Differences between work methods and gender in computer mouse use. *Scandinavian Journal of Work and Environmental Health, 26*(5), 390-397.

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.

Wang, Y., Peng, P., Liang, L. J., Zhang, W., and Wu, S., (2007). Road hazard reaction testing using driving simulation: the novice vs. the experienced drivers. In M. Helander, M. Xie, M. Jaio, K. C. Tan (Ed.s), *2007 IEEE International Conference on Industrial Engineering and Engineering Management, Vols 1-4,* Singapore.

Watts, G.R. and 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. Department of Transport, DOT. London, UK.

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

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 and Prevention, 42*(4), 1232–1239. http://dx.doi.org/10.1016/j.aap.2010.01.017

Williams, A. F., & Tefft, B. C. (2014). Characteristics of teens-with-teens fatal crashes in the United States, 2005–2010. *Journal of Safety Research, 48,* 37-42.

Wolfe, J. M. (1994). Guided search 2.0: a revised model of visual search. *Psychonomic Bulletin and Review, 1*, 202-238.

World Health Organisation (2015). Global status report on road safety, 2015. http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/ (last accessed on 19/12/16).

Yamani, Y., Samuel, S., Knodler, M.A., Fisher, D.L. (2016). Evaluation of the effectiveness of a multi-skill program for training younger drivers on higher cognitive skills. *Applied Ergonomics, 52,* 135-141.

Yamauchi, T., Seo, J. H., Jett, N., Parks, G., and Bowman, C. (2015). Gender differences in mouse and cursor movements. *International Journal of Human-Computer Interaction, 31*(12), 911-921.

Yeung, J.S., and Wong, Y.D. (2015). Effects of driver age and experience in abrupt-onset hazards. *Accident Analysis and Prevention, 78,* 110-117.

Zhai, S. M., Kong, J., Ren, X. S. (2004). Speed-accuracy tradeoff in Fitts' law tasks – on the equivalency of actual and nominal pointing precision. *International Journal of Human-Computer Studies, 61*(6), 823-856.