Cross-cultural effects on drivers' hazard perception: Validating a test paradigm for developing countries

Phui Cheng Lim

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

December 2015

ABSTRACT

The hazard perception skill of a driver refers to their ability to identify potentially dangerous events on the road, and is one of the only driving-specific skills that has been consistently linked to accident rates. Hazard perception tests are used in several developed countries as part of the driver licensing curriculum, however little research has been done in developing countries where road safety is a primary concern. The extent to which hazard perception skill transfers to different driving environments is also unclear. This thesis therefore has two major aims: to examine hazard perception in a cross-cultural context, and to validate a hazard perception test for potential use in driver licensing in lower-income, developing countries.

Most of the experiments in this thesis compare hazard perception skill in drivers from the UK – where hazard perception testing is well established – and drivers from Malaysia – a developing country with a high accident rate where hazards frequently occur. Typically, hazard perception skill is assessed by showing participants clips filmed on the road and asking them to respond as soon as they detect a developing hazard, with shorter response times reflecting greater levels of skill. Chapter 2 presents evidence that Malaysian drivers may be desensitized to hazardous road situations and thus have increased response times to hazards, creating validity issues with the typical paradigm. Subsequent chapters therefore use a predictive paradigm called the "What Happens Next?" test that requires drivers to predict hazards, leaving performance unaffected by hazard desensitization. Malaysian drivers predicted hazards less accurately than UK drivers in all cross-cultural experiments,

indicating that exposure to a greater number of hazards on Malaysian roads did not have a positive effect on participants' predictive hazard perception skill. Further experiments indicated that explicit knowledge plays a minor role in the "What Happens Next?" test, and that experienced drivers appear to compensate for reduced visual information more effectively than novices. Experienced drivers from both Malaysia and the UK also outscored novices in all experiments using the predictive paradigm, suggesting the "What Happens Next?" test provides a valid measure of hazard perception skill and may offer a practical alternative for hazard perception testing in developing and even developed countries.

ACKNOWLEDGEMENTS

Throughout my PhD, Lizzy Sheppard has been a model of leadership and professionalism, and I am extremely grateful to have benefitted from her supervision. Lizzy, a huge thank you for your guidance and support, your unfailing encouragement and positivity, and for giving me the freedom to find my own way in my research. I couldn't have asked for a better supervisor. I also owe an immense debt of gratitude to Matt Johnson, for being an endless source of knowledge, advice, and geekery – and whose inability to answer any question in less than half an hour gave rise to a great deal of learning and inspiration over the last four years. Thank you for being infinitely nerdsnipable. Thanks also to David Crundall and Peter Chapman for supporting my work on the UK campus, and for the feedback, advice, and thought-provoking conversations. Neil Mennie and Mark Horswill definitely deserve a special mention for (hopefully) reading all 200-some pages of this thesis, and thanks goes to David Keeble for providing various support in the late stages of my PhD. So many of my participants made me appreciate the kindness of strangers – 99% of them will never read this, but thank you all the same for waking up early, travelling into town, and recruiting small armies of friends for my experiments. I couldn't have written this thesis without all of you.

The Nottingham psych community has been very supportive, and I am lucky to have spent time on both the UK and Malaysian campuses and gotten to know the departments there. Thanks to the UK folks for always taking the time to talk shop during my visits and including me in pub rounds (often the same thing). In Malaysia,

Ian Stephen has always had my back for work and advice, and Jess Price has done the same PLUS baked me things. And of course, to *all* those who came through BB47 at some point, my Permanent Head Damage folks, I would certainly not have survived the last four years without you! Thanks for the laughs, the commiseration, the general ridiculousness, and so much more.

Outside of Nottingham, I am especially grateful to Ross Bauer for his constant support and friendship. Russell and Camille Boyd, Andrew Fernandes, Steph Kong, Sook-Chen Lee, and Sarah Quek have also provided invaluable humor and encouragement and generally been the best of friends. And finally, I would like to thank my family for their love and support throughout my PhD – especially my mother, who I think is still slightly disappointed that my work doesn't allow me to read minds, but is very proud of me anyway.

TABLE OF CONTENTS

LIST OF FIGURES

All error bars in figures represent standard error of the mean.

LIST OF TABLES

LIST OF ABBREVIATIONS

r_{bis}	Biserial correlation coefficient
BP	Behavioral prediction (Crundall et al., 2012)
DF	Divide and focusing (Crundall et al., 2012)
EP	Environmental prediction (Crundall et al., 2012)
FDR	False discovery rate
HP	Hazard perception
HPT	Hazard perception test
MY	Malaysia
QT-HPT	Queensland Transport Hazard Perception Test

CHAPTER 1

GENERAL INTRODUCTION

1.1 Hazard perception

In 2010, the United Nations General Assembly announced the 2011 – 2020 period as the Decade of Action for Road Safety, calling for a move to stabilize and reduce the rising number of road traffic accidents globally. The World Health Organization's subsequent Global Status Report on Road Safety cited road injuries as the eighth leading cause of death worldwide in 2012, the only one of the top ten unrelated to health, and estimated that by 2030, it will become the fifth leading cause, barring major interventions (Toroyan, 2013).

One point noted in the WHO report was the impact of road traffic accidents on young people aged 15 – 29, for whom road injuries were the leading cause of death. While many factors affect road fatalities – for instance speed, drink driving, helmets, seat belts, and child restraints (Toroyan, 2009) – it is widely documented that young, newly licensed drivers contribute disproportionately to road accidents (Deery, 1999; Mayhew, Simpson, & Pak, 2003; McKnight & McKnight, 2003). Generally, accident rates are highest for drivers immediately after licensing, after which a steep initial decline occurs. While age and driving experience are often confounded, research suggests that experience plays a major role in declining accident rates (McKnight & McKnight, 2003).

While several skills develop with driving experience, not all of them influence accident involvement within the population. According to Horswill and McKenna (2004), among driving-specific skills, only one has been noted to correlate with crash risk across multiple studies: hazard perception, or the ability to identify potentially dangerous situations on the road. Pelz and Krupat (1974) presented participants with films of driving scenes, and instructed them to imagine they were the driver in the film scenarios and to move a lever on a sliding scale from "safe" to "unsafe" according to the level of danger they felt was present. They calculated participants' response times to hazards in the scenes, and found that non-accident-involved participants recorded faster response times compared to those who had been involved in traffic violations or accidents. Since its conception, hazard perception has been assessed in a variety of other ways. For instance, rating the level of risk present in driving scenes (Finn & Bragg, 1986; Groeger & Chapman, 1996; Pelz & Krupat, 1974; Wallis & Horswill, 2007), analyzing drivers' eye movements while watching videos or driving (Chapman & Underwood, 1998; Mourant & Rockwell, 1972), predicting hazards about to occur (Crundall, 2016; Vlakveld, 2014), or describing hazards encountered during on-road drives (Soliday, 1974).

However, reaction time-based hazard perception tests remain common, where participants are often explicitly instructed to respond to a hazard as fast as possible. Several methods are used to record these responses; for instance, moving a lever (Crundall, Chapman, Phelps, & Underwood, 2003; Pelz & Krupat, 1974), using a touchscreen (Horswill, Anstey, Hatherly, & Wood, 2010; Scialfa et al., 2011; Wetton et al., 2010), clicking on them with a mouse (Wetton, Hill, & Horswill, 2011, 2013), or simply pressing a response button (Sagberg & Bjørnskau, 2006; Underwood, Phelps, Wright, Van Loon, & Galpin, 2005; Wallis & Horswill, 2007). In all these

cases, shorter response times are thought to reflect greater levels of hazard perception skill, and there has been some evidence that this measure of performance is associated with lower crash risk (see Section 1.1.2; (Boufous, Ivers, Senserrick, & Stevenson, 2011; Horswill et al., 2010; Horswill, Hill, & Wetton, 2015; Pelz & Krupat, 1974; Quimby, Maycock, Carter, Dixon, & Wall, 1986; Wells, Tong, Sexton, Grayson, & Jones, 2008). Drivers with more post-licensing experience also tend to outperform newly qualified drivers, with driving experience considered a proxy for safety, given that crash risk tends to decrease as experience accrues (Horswill & McKenna, 2004) for a review; Sections 1.1.2 and 1.1.3 for more detail).

Partly due to the evidence supporting a link between hazard perception and crash risk, several countries have incorporated some form of computer-based hazard perception testing into their driver licensing process, including the UK, the Netherlands, and Australia (see Section 1.1.1 for more detail on licensing in Australia and the UK). In fact, McKenna and Crick (1997) argue that a laboratory setting possibly allows more accurate assessment of hazard perception skill, given the greater degree of control and variety of hazard encounters allowed over an on-road setting. Watts and Quimby (1979) presented early evidence that simulator studies produce similar results to on-road driving, and later research confirmed that a fully instrumented simulator, on-road driving, and watching video clips all yielded comparable patterns of behavior (Underwood, Crundall, & Chapman, 2011).

The current chapter will review the role of reaction time-based hazard perception tests in driver licensing (Section 1.1.1), and the links between performance on these tests and later accident involvement (Section 1.1.2). Since it is not always practical to conduct the large-scale studies required to accurately gauge crash risk, much hazard perception work uses driving experience as a proxy for safety, although

some studies have failed to find a link between experience and response latency, possibly due to the types of hazards used (Section 1.1.3). Several alternative tasks to the reaction time test have therefore been proposed, although none have yet shown direct links to accident rates (Section 1.1.4). Past research has also found that experienced drivers' visual strategies significantly differ from novices', possibly contributing to their superior hazard perception skill (Section 1.1.6). However, hazard perception skill and to a certain extent visual strategies can be improved via various training methods (Section 1.1.5). Finally, we will discuss the lack of hazard perception research in developing countries, where road safety is a major concern (Section 1.2).

1.1.1 In driver licensing

Some form of hazard perception assessment is used as part of the licensing process in the UK, Australia, and more recently, the Netherlands. A hazard perception test, based on the button press response latency measurement described above, was introduced in the UK licensing process in November 2002. Prospective UK drivers are first required to hold a provisional license, which has a minimum age limit, but no driving-specific requirements. They then must pass a theory test, which consists of a multiple choice test and a hazard perception test. After passing the theory test, they may sit for the on-road practical test, which allows them to hold a full, unrestricted license.

The hazard perception test in the UK consists of 14 videos and 15 hazards (UK Department for Transport, 2016), and previously used filmed clips, which were updated to computer-generated imagery in January 2015. Drivers can score a maximum of 5 points on each hazard depending on how early in its window they

respond, and must score at least 44 of a possible 75 points (59%) to pass the test. For driving instructors, the pass mark increases to 57 of 75 (76%).

In Australia, the driver licensing process differs between states, but all have at least three stages: a learner's license (the equivalent of the UK's provisional license), a provisional license, and finally, a full, unrestricted license, with learner and provisional stages sometimes split into two further stages. Of the eight states, five require a video-based hazard perception test at some point in the licensing process, although one of the remaining three, Tasmania, requires a hazard commentary as part of the on-road test. While the stage at which drivers sit a hazard perception test varies, it is always required to obtain either a first or second stage provisional license, which must be held for at least two years before progressing to a full license. The nature of the hazard perception test also varies depending on the state. For instance, New South Wales, South Australia, Victoria, and Western Australia incorporate situations that call for a variety of driving maneuvers such as slowing down, overtaking, or taking a turn, and test-takers indicate appropriate times to commence these actions. New South Wales also requires drivers to take a second, more advanced hazard perception test as the last step in obtaining a full license. Some researchers however argue that including a wider variety of actions might invoke multiple constructs rather than hazard perception skill alone, since many of these judgments involve some form of risk-taking (Horswill, Hill, et al., 2015; Wetton et al., 2011). The Queensland Hazard Perception Test (QT-HPT) on the other hand follows the traditional research paradigm described above more closely, and requires test-takers to click on developing hazards (for more detail on this study and creation of the QT-HPT, see Section 1.1.3; also Wetton et al., 2011).

1.1.2 Accident involvement

Several studies have linked hazard perception test performance to accident rates, either directly or indirectly. Pelz and Krupat (1974) found that participants with clean safety records during the past year recorded faster response times to hazards, compared to those involved in traffic violations or accidents respectively. Horswill et al. (2010) also reported a link between response latency and crash involvement in drivers aged 65 and over. However, as Horswill, Hill, et al. (2015) note, retrospective studies allow the possibility that a driver's hazard perception skill was altered as a result of their accident involvement, rather than hazard perception skill, or a lack of, being the causal factor. Furthermore, given the low frequency of accidents, it is difficult to definitively conclude a link between hazard perception and accident involvement without very large participant numbers (Horswill & McKenna, 2004). Given this, several large-scale studies have linked performance on the hazard perception licensing test to subsequent accident rates, in both the UK and Australia.

Wells et al. (2008) surveyed drivers who had taken the practical licensing exam in the UK from 2001 to 2005; because the hazard perception component was introduced as part of the theory test in 2002, their data allowed them to compare accident involvement in respondents who had and had not taken the hazard perception test. While univariate analyses were inconclusive, subsequent multivariate analyses suggested that the hazard perception component had had a positive effect on road safety, even controlling for age, gender, driving experience, and exposure to different driving environments. In other words, drivers who had taken the hazard perception test had slightly lower liability for some types of accidents in their first year of driving, compared to drivers who had not. Wells et al. estimated that drivers who had taken the hazard perception test showed a reduction in accident liability for non-low-

speed public road accidents where the driver accepted some blame (at least 3%), and all non-low-speed public road accidents (reduction of at least 0.3%). They also compared the highest and lowest scoring groups on the hazard perception test, and estimated that for non-low-speed public road accidents where the driver accepted some blame, the highest scoring group's accident liability was at least 4.5% lower.

In Australia, Boufous et al. (2011) examined the crash rates of newly licensed drivers in New South Wales in relation to their on-road and hazard perception test results during the licensing process, with the number of failures on each test being used as the primary variables. They obtained police crash records for study participants, a contrast from the other studies in this section that use self-reported data. Accident data was retrieved approximately two years after participants had passed the on-road test, which was also approximately one year after passing the hazard perception test (since New South Wales licensing procedure requires drivers to wait at least one year after the on-road test before they can take the hazard perception test). While failing the on-road test did not correlate with failing the hazard perception test, crash risk was linked to both these tests. After controlling for sociodemographic and behavioral factors, Boufous et al. reported that failing the on-road test at least 4 times was associated with a significantly higher later crash rate, compared to passing the test on the first attempt. Similarly, failing the hazard perception test at least twice was associated with a significantly higher later crash rate. Interestingly, these effects seemed to play out differently depending on gender. For the on-road test, slightly more females than males passed on their first attempt, but the crash risk of those who failed the test at least 4 times was especially high for females. In contrast, for the hazard perception test, slightly more males than females passed on their first attempt, but failing the test at least twice was associated with

significantly higher crash rates for males, while it was not for females. It should also be noted that the hazard perception test administered in New South Wales differs slightly from the ones used in the UK and Queensland (for more detail, see Section 1.1.1), but nevertheless the general results from all these studies appear similar.

Horswill, Hill, et al. (2015) conducted a study on drivers taking the QT-HPT from 2008 to 2012. One key difference between this test and the UK hazard perception test is that it requires test-takers to identify the source of the hazard in question by clicking on it with a mouse, rather than simply pressing a button to indicate a hazard has occurred; the Queensland test therefore elicits responses for both time and location, rather than only time. A similar version of this paradigm, where drivers used a touchscreen to identify hazards, has previously been linked to crash risk in older drivers (Horswill et al., 2010). Immediately after taking the QT-HPT, test-takers were asked if they would like to participate in a survey; respondents were then taken to an online survey where they then provided their hazard perception score along with other information such as demographics and driving experience. They received a second follow up survey one year later. After controlling for age, gender, and driving experience, Horswill, Hill, et al. found that respondents' HPT score was linked to both retrospective (at the time of test-taking) and prospective crashes (one year after test-taking). This was true for all active crashes, where a one second increase in hazard perception reaction time corresponded to a 10% increase in the chance of being involved in a retrospective crash, and a 21% increase for a prospective crash. Hazard perception score also seemed to significantly affect the chances of being involved in several sub-types of active crashes, namely ones involving failures of attention and anticipation.

1.1.3 Driving experience

1.1.3.1 *As a proxy for safety*

Road traffic deaths are the usual indicator by which overall road safety is measured, and are relatively well documented compared to for instance, accidents or non-fatal accident injuries. The latter two metrics are often unreported and/or poorly documented, especially in lower-income countries. For instance, many accidents go unreported and thus no official record of them exists. Police recording may also be inconsistent and/or inaccurate when it does happen, especially when reporting injuries, as police often lack the requisite training to correctly categorize injuries (Toroyan, 2013).

However, for smaller-scale research studies, it is often impractical to use road accidents or fatalities as a metric (Horswill & McKenna, 2004). Accidents are by nature infrequent and complex, affected by many factors beyond a driver's control; as a result, in a study with a small participant pool, the difference in hazard perception skill between a driver with three minor accidents and a driver with none might actually be negligible. As a result, the statistical power required to gain an accurate insight into accident/fatality rates and driver safety is extremely high.

Because of this, the majority of hazard perception studies use driving experience rather than accident or fatality rates; since the accident propensity of novice drivers is considerably higher than experienced drivers, experience is considered an acceptable proxy for driver safety. Driving experience is usually operationalized as licensing time and/or mileage. Although age and experience are often confounded, research has shown accident rates to decrease with both; McKnight and McKnight (2003) and Horswill and McKenna (2004) argue that experience is crucial to reduced accident rates regardless of age.

1.1.3.2 *Experience and response latency*

Various previous studies support the notion that experienced drivers respond faster to hazards than novices. For instance, Wetton et al. (2011) describe creating a hazard perception test for use in the driver licensing exam in Queensland, Australia (the QT-HPT), and propose five principles for test creation; among them, they stress that a licensing test must meet higher standards than a test being used for research. Their test stimuli were therefore evaluated by driving instructors from Queensland Transport and the types of hazards contained therein reflected the types of crashes that most commonly occur among novice drivers, resulting in 91 items in total. They also conducted a brief validation study to confirm that the test instructions were accessible to participants with a low level of English skill, aiming for a reading age equivalent of a 10-year-old native English speaker. In order to minimize inappropriate and/or ambiguous responses, participants were asked to click on the relevant road users with the mouse and informed that clicking on non-relevant road users too many times would result in failing the test. Wetton et al. reported that in the full 91-item test, experienced drivers showed significantly faster response latencies and a higher response rate than learners. Additional tests for simple reaction time and mouse skills revealed that learner drivers were faster to respond in both tests. To create the final version of the test used in the QT-HPT, Wetton et al. ranked each clip within its crash-type category according to the effect size of the experience difference (measured by Cohen's *d*), and generated four 15-clip tests, with each individual test retaining high reliability and a significant learner/experience difference.

Scialfa et al. (2011) administered a hazard perception test to novice and experienced drivers in Canada, providing participants with a touchscreen to record the location as well as time of their hazard responses. After removing videos where even

experienced drivers had insufficient hit rates and those with multiple hazard onsets occurring simultaneously, leaving 49 of 64 videos, they reported that novices missed significantly more hazards than experienced drivers and had slower response times, although there was no difference in false alarm rates between the two groups in videos that did not contain hazards. Scialfa et al. also assessed their test's discriminant validity based on several different variables, using binary logistic regressions. Two initial analyses to predict driver group membership were conducted with first, reaction time alone and second, reaction time, miss rates, and false alarm rates, with mixed results. All scenes were then removed for the reaction time and miss rate measures except for those where experienced drives had significantly outperformed novices, leaving 18 videos from the earlier 49. Classification accuracy was similarly high and novice classification accuracy was slightly higher. Scialfa et al. reported that the final 18-item test had good reliability and a significant novice/experience difference for reaction time.

Horswill et al. (2008) reported a two-experiment hazard perception study with novice and experienced drivers, and later older drivers. In the first experiment, experienced drivers had significantly faster response times than novices, and there was no difference in response rate, which was near ceiling, between driver groups. After removing scenes that failed to differentiate experience and adding several new scenes, they administered the modified hazard perception test to older drivers (>65 years of age), along with several cognitive and visual measures. Horswill et al. reported that while increased age correlated with declining hazard perception performance, useful field of view and contrast sensitivity also similarly declined. Further analyses revealed that contrast sensitivity, useful field of view, and simple reaction time were the three primary drivers of hazard perception performance, and

age and visual acuity had a minimal impact. Their results suggested that while hazard perception skill does appear to decline with age, cognitive and visual skill may account for much of the individual differences in older drivers.

1.1.3.3 *Response latency inconsistencies*

While many studies have found that experienced drivers respond to hazards faster than novices (Horswill et al., 2008; Scialfa et al., 2011; Wallis & Horswill, 2007; Wetton et al., 2011), several others report no hazard perception latency differences regardless of driver experience. For instance, Chapman and Underwood (1998) found differences in several eye tracking measures between novice and experienced drivers, but no difference in response latency (for a more detailed description of this study, see Section 1.1.6). Similarly, Crundall et al. (2003) reported no differences in hazard ratings regardless of driving experience, but found oculomotor and physiological differences. They recruited three groups of drivers; a novice driver group, a police driver group, and a control group that was age- and experience- matched with the police drivers. Participants watched 48 videos, half filmed during the daytime and half at night; those filmed from police vehicles engaged in pursuit of other vehicles, police vehicles engaging in emergency-response driving, and standard control drives. Their responses were measured using a sliding scale, where they constantly adjusted a slider to reflect the hazardousness of the current situation. Their eye movements and electrodermal responses (EDRs) were also recorded. Crundall et al. observed no differences in either hazard ratings or number of hazards reported regardless of driving experience, but found several differences in fixation duration, spread of search, and EDRs. Novices had the longest fixation durations for all clips, and showed the greatest effect of attentional narrowing (for more detail on attentional narrowing, see Chapman & Underwood 1998; Section

1.1.6); for instance, while all drivers showed increased fixation durations during nighttime pursuit clips, this was especially true for novices. Novices also had a smaller horizontal spread of search than the control and police groups, although this was primarily driven by the police drivers having very wide horizontal scanning. Finally, they reported no differences in mean EDR across clips for driver experience, but police drivers produced a higher number of discrete EDRs compared to matched controls, possibly suggesting that on a physiological level, they were responding to more potentially hazardous events. EDRs were also higher and more frequent during daytime clips for all drivers, which Crundall et al. attributed to a greater amount of visual information during the day.

Sagberg and Bjørnskau (2006) administered a hazard perception test containing 31 scenes to four groups of drivers. Three of these groups had held a Norwegian driving license for 1, 5, and 9 months and were considered novice drivers. The fourth group, composed of experienced drivers, had held their licenses for an average of 27.1 years. During one of the two experimental blocks, drivers also completed a secondary task which required them to calculate the total of several one-digit numbers that were read every 10 seconds throughout the block. Sagberg and Bjørnskau analyzed response rates and response times for each individual hazard, and overall response times for all hazards, replacing missing responses with the maximum possible reaction time. They found no overall differences in response rate or response latency between the three novice driver groups, nor between the novice and experienced driver groups after collapsing the novice driver data. They observed some differences in individual hazards; in one of the 31 hazards, the least experienced group (the 1-month novices) had a significantly lower response rate than the other two novice groups. In the novice and experienced driver comparison, novices had

significantly lower response rates for two hazards, and significantly longer reaction times for six hazards (including the two with lower novice response rates). Finally, novice male drivers had longer reaction times with the secondary task compared to both novice female and experienced male drivers, but there were no other differences in secondary task performance.

Underwood, Ngai, and Underwood (2013) also reported no response latency differences between novice and experienced drivers, although a third group of experienced drivers who also rode motorcycles (rider-drivers) showed faster response latencies. Participants viewed videos of hazards that were classified as either gradual onset or abrupt hazards, depending on how long the involved road users were on screen before an avoidance or braking response became necessary; hazards where the road users were on screen for more than 3 seconds were classified as gradual onset, while less than 3 seconds were considered abrupt hazards. For abrupt hazards all participants had similar response times, but for gradual onset hazards, rider-drivers were faster to respond than both the novice and experienced driver groups, while there was no difference between the two latter groups. Participants also responded to abrupt hazards faster than gradual onset hazards. There was no difference between any of the driver groups for the number of hazards detected, although all participants detected more gradual onset than abrupt hazards. Hazard selectivity was also calculated using a ratio of hits to false alarms, where experienced drivers showed greater selectivity than rider-drivers (i.e. fewer false alarms compared to hits), and all participants had greater selectivity in videos with gradual onset hazards than abrupt (i.e. fewer false alarms compared to hits), although this should not be interpreted as greater selectivity to gradual onset hazards per se, since hazards other than the pre-defined ones were not classified.

1.1.3.4 *Anticipatory cues and precursors*

Some researchers have suggested that the inconsistency in experience differentiation may stem in part from the particular types of hazards used in these studies. Sagberg and Bjørnskau (2006) noted that the elements of complexity, surprise and anticipation were all important in the videos that did differentiate experience, although they noted that other, non-differentiating videos also shared one or more of these characteristics. Wetton et al. (2011) suggest five principles for effective test creation that they used in their development of the QT-HPT, one of them being that "a hazard perception test should discriminate between individuals on the basis of differences in hazard perception skill, not differences in simple response time" (p. 1760). They discuss the importance of choosing scenes where predictive cues to the hazard are present, and argue that scenes without anticipatory cues tend to test simple reaction time rather hazard perception skill. Predictive cues, then, become especially important in order for drivers to demonstrate good hazard perception skill, especially as younger, inexperienced drivers tend to have faster simple reaction times than older, experienced drivers and so might even outperform them in abrupt-onset hazards. This certainly seems to be the case in Underwood et al. (2013) where the only difference in response latency between driver groups was found in gradual onset hazards and not abrupt hazards, although this difference was found in the rider-driver group rather than between novice and experienced drivers.

Garay, Fisher, and Hancock (2004) conducted a simulator study with novice and experienced drivers while recording their eye movements. Participants drove two simulated routes, one simulating daytime and one simulating nighttime. Each route contained nine scenarios that fell into three different categories: pedestrians, traffic control devices, and conflicting traffic. Each scenario had an advance cue, critical

area and critical element. For example, in one pedestrian scenario, a truck is parked in a residential area just in front of a crosswalk, potentially hiding a pedestrian about to enter said crosswalk. The critical area is therefore the area in front of the truck, where a pedestrian might be standing. While the truck hides the sign indicating the crosswalk itself, drivers should still be aware of the crosswalk ahead due to a warning sign (advance cue). The critical element in this case was a pedestrian crossing the road several seconds before the driver reached the crosswalk; this was intentionally timed so the pedestrian would not be a potential hazard to the driver, but instead served to remind drivers that another pedestrian might also step into the road while obscured by the truck. Participants were judged to have recognized the risk the scenario posed if they looked at the critical area in front of the truck while passing it, where a pedestrian might be standing. Garay et al. reported that all drivers recognized risks more often in the daytime, and experienced drivers recognized risks more often than novices did. This was true in the traffic control devices and conflicting traffic categories; however, in the pedestrian category, novices recognized risk as often as experienced drivers did, possibly due to their own experience as pedestrians. Finally, while the presence of the foreshadowing elements did appear to aid all drivers in recognizing potential risks, it seemed to benefit experienced drivers more than novices, since they were more likely than novices to fixate the critical area given they had first fixated the critical element. Similarly, novices were also more likely to miss the critical area despite fixating the critical element than experienced drivers were. These results suggest that while foreshadowing elements benefit all drivers, novice drivers lack the experience to make full use of them.

In another simulator study, Crundall et al. (2012) examined the responses of learner drivers, experienced drivers and driving instructors to different types of

hazards, measuring approach speed and fixations on critical stimuli (i.e. hazards and precursors). They defined three types of hazards: behavioral prediction (BP) hazards, where a hazard and its precursor are the same road users (for instance, a pedestrian walking beside the street – precursor – who then steps into the road – hazard); environmental prediction (EP) hazards, where the hazard and precursor are two different elements (an ice cream van parked on the roadside – precursor –, from behind which a child steps into the road – hazard); and divide and focusing attention (DF) hazards, where multiple precursors exist, one of which eventually develops into a hazard (a parked bus – EP precursor – and a pedestrian – BP precursor – on opposite sides of the road, where the pedestrian crosses the road to the bus – hazard). Participants drove a simulated route through a virtual city where nine hazards, three of each type, were triggered at various points. Crundall et al. reported that learner drivers fixated fewer BP precursors and EP hazards than both experienced drivers and instructors, but there was no difference in BP hazards and EP precursors; furthermore, all drivers fixated more hazards than precursors. Learners were also the slowest of all driver groups to fixate BP precursors, BP hazards, EP hazards, and DF hazards (in fact, almost all critical stimuli except for EP and DF precursors). Dwell time analyses also revealed that all drivers spent more time fixating on hazards than precursors, and experienced drivers fixated critical stimuli longer than instructors, who in turn fixated critical stimuli longer than learners. Finally, driving instructors showed different speed signatures when approaching BP and DF events, decreasing their speed more than the other driver groups.

The above studies indicate that, as Wetton et al. (2011) suggest, driving experience plays a key role in whether drivers recognize anticipatory cues, and different types of cues may also be viewed differently by drivers with varying

experience. It seems possible that different types of hazards in a hazard perception test may differentiate experience more or less effectively, which recent research may indicate (Crundall, 2016; Section 1.1.4).

1.1.4 Alternative tasks

Given the potential issues with a reaction time-based paradigm, several researchers have developed alternative paradigms to measure hazard perception skill (Castro et al., 2014; Crundall, 2009, 2016; Huestegge, Skottke, Anders, Müsseler, & Debus, 2010; Jackson, Chapman, & Crundall, 2009; Scialfa et al., 2012, 2012; Scialfa, Borkenhagen, Lyon, & Deschênes, 2013; Vlakveld, 2014; Wetton et al., 2010). Some of these tasks involve responses that are unaffected by drivers' judgments of hazardousness, arguably decreasing subjectivity as one can be reasonably certain these judgments are not confounding responses (Castro et al., 2014; Crundall, 2009, 2016; Jackson et al., 2009; Wetton et al., 2010). Others employ static rather than dynamic stimuli, which also decreases subjectivity and offers practical advantages such as reduced time spent for both test preparation and administration (Crundall, 2009; Huestegge et al., 2010; Scialfa et al., 2012, 2013; Wetton et al., 2010). While none of these tasks are inherently superior to others (apart from how well they differentiate driving experience), they serve to highlight some of the issues with the traditional task that researchers feel are particularly important to address. Furthermore, while other assessment methods certainly exist, such as ratings of risk (Groeger & Chapman, 1996; Pelz & Krupat, 1974), some of these provide fairly subjective measures of performance and in a non-research setting, can be easily deceived by test-takers simply applying higher levels of caution than they otherwise would. We will primarily focus on tasks that could be conceivably used in a licensing procedure.

Wetton et al. (2010) suggest three components of hazard perception skill as measured by the traditional task:

> (1) drivers must register the existence of the potentially hazardous event (this step is defined as *hazard detection* within [their] paper);
>
> (2) drivers must then make a judgement regarding whether the trajectory of either the potential hazard and/or their own car has the potential to cause a conflict;
>
> (3) drivers must then classify the event as to whether it warrants a response (defined as *hazard classification* within [their] paper) (p. 1232).

We shall use Wetton et al.'s terms for the first and third components (*hazard detection* and *hazard classification*, respectively), and refer to the second as *trajectory judgment*. Wetton et al. argue that in the reaction time paradigm, each of these components to a certain extent depends on the drivers' judgment of the prior one. Wallis and Horswill (2007) conducted a signal detection experiment that suggested novices' slower hazard response times may stem from having different hazard classifications to experienced drivers; in other words, a given hazard must be more hazardous before it passes their threshold for responding. Because of this, several alternative tasks have focused on one or two of these components rather than all three, although this has not always resulted in experience differentiation (Wetton et al., 2010).

We will therefore discuss three general categories of alternative tasks. First, those which incorporate only the first component, hazard detection, often measured using change detection tasks (Groff & Chaparro, 2003; Wetton et al., 2010). Second, tasks that emphasize anticipatory skill, often asking drivers to predict hazards that

have not yet materialized, effectively eliminating the third component of hazard classification (Castro et al., 2014; Crundall, 2016; Jackson et al., 2009; Vlakveld, 2014). Finally, tasks that use static rather than dynamic stimuli (Huestegge et al., 2010; Scialfa et al., 2012, 2013; Wetton et al., 2010).

1.1.4.1 *Change detection*

Wetton et al. (2010) conducted two experiments using a change detection task and traditional hazard perception tests (HPTs). All drivers completed three hazard perception tests, using footage from the Australian Central Territory (ACT), Queensland (QLD), and the UK; the latter two tests had been previously validated with Australian and UK drivers respectively. They also completed a change detection task which used images from only the ACT footage. Drivers viewed two images of driving scenes that were identical, save for the presence or absence of a potential hazard, which they had to detect as quickly as possible. In the first experiment, Wetton et al. reported that experienced drivers responded faster than novices in the ACT HPT, and response times for all three hazard perception tests were highly correlated. However, reaction time for the change detection task did not correlate with any of the three hazard perception tests; furthermore, novices responded faster than experienced drivers, suggesting performance was not correlated with crash risk. In contrast, in a second experiment with older drivers aged 65 and over, they found significant positive correlations between reaction time for the ACT HPT and the change detection task. They also found that as age increased and useful field of view (UFOV) decreased, reaction times for both tasks also increased. Notably, however, the QLD HPT correlated with the ACT HPT but not the change detection task. Wetton et al. tentatively concluded that the change detection task may be appropriate

in older drivers, although it is possible that it may measure a general ability rather than hazard detection itself.

Groff and Chaparro (2003) conducted a change detection task with non-drivers, experienced drivers, and police drivers. Similar to Wetton et al. (2010) above, participants viewed two images of driving scenes presented alternately, identical except for a change target. Unlike Wetton et al. (2010) however, the change targets were not necessarily potential hazards, but were vehicles, traffic signs, or other objects in the environment such as buildings, trees, billboards etc. Changes to vehicles and traffic signs were considered task-relevant, and changes to objects in the environment were task non-relevant. There were three further types of potential changes: presence, where the change target was present or absent (the same and only type of change used by Wetton et al.); position, where the change target moved to a new location; and feature (where some feature of the change target was modified, for instance color or size). Groff and Chaparro found that all participants detected changes to traffic signs fastest, then vehicles, then objects in the environment, and in fact all participant groups detected task-relevant changes faster than task non-relevant. Driving experience also played a key role in detection speed: police drivers detected vehicle and traffic sign changes faster than experienced drivers, who in turn detected these changes faster than non-drivers. Finally, the type of change also affected detection times for vehicles and objects in the environment, but not traffic signs. For vehicles, participants detected presence and position changes faster than feature changes, while for objects, participants detected feature changes faster than both presence and position changes. Notably, this experiment was conducted as a change detection task using driving stimuli, rather than a test intended to differentiate driving experience. However, it is interesting that driving experience played a role in

detection speed, particularly considering this was not the case in Wetton et al. (2010)'s change detection task which used actual potential hazards. While this could stem from driving experience being more disparate, it may also be because of the different nature of the changes, or methodological differences such as in timing or the use of a touchscreen in Wetton et al. which may have given a disproportionate advantage to younger drivers. It should be noted however that Galpin, Underwood, and Crundall (2009) also found no effects of experience in another change detection task involving driving scenes.

1.1.4.2 *Anticipatory skill*

Vlakveld (2014) describes two tasks where learner drivers and professional drivers (driving instructors and examiners) had to detect overt and covert latent hazards, analogous to Crundall et al. (2012)'s behavioral prediction (BP) and environmental prediction (EP) hazards respectively (see Section 1.1.3 for more detail). In Task 1, participants first watched computer-generated clips with no particular task other than general vigilance for hazards, and then re-watched the clips and identified the point in time and location at which they felt danger was most imminent. In Task 2, a different set of participants completed a hazard perception task with the same clips. After each clip, all screenshots taken at the time point of hazard detection were displayed, and participants chose the screenshot they felt contained the highest priority hazard. Professional drivers significantly outscored learners on both tasks, although this difference was smaller in Task 2. Furthermore, in Task 1, professional drivers outscored learners in detecting both BP and EP hazards, although all drivers had higher scores for BP than EP hazards. This was not the case in Task 2, where only learners had higher scores for BP hazards; furthermore, professionals did not outscore novices on these hazards. Interestingly, learners who

played computer games also outscored learners who did not, but only in Task 2. Vlakveld concluded that Task 1 was preferable to Task 2.

Jackson et al. (2009) explored a predictive paradigm called the "What Happens Next?" test, which traditionally has primarily been researched as a driver training tool (Chapman, Underwood, & Roberts, 2002; Horswill, Falconer, Pachana, Wetton, & Hill, 2015; Horswill, Taylor, Newnam, Wetton, & Hill, 2013; Wetton et al., 2013), but has also since been employed as a potential diagnostic (Castro et al., 2014; Lim, Sheppard, & Crundall, 2014). Novice and experienced UK drivers watched clips filmed on the road, but unlike the traditional hazard perception test, these clips were stopped in the middle, usually immediately before hazard onset. Crucially, at this stopping point, the clip contained enough information for an intelligent observer to correctly predict what would happen next. Drivers then answered three questions: (1) What was the hazard? (2) Where was the hazard? (3) What happens next? Two versions of this test were used: a cut to black condition, where the screen immediately cut to black after the stopping point, and a freeze frame condition, where the final still image remained on screen for a further 20 seconds. Jackson et al. reported an interaction of condition and driver experience, which was driven by novice drivers' lower scores in the cut to black condition, both compared to their scores in the freeze frame condition and experienced drivers' scores in the cut to black condition. They also found that for all drivers, scores dropped significantly with each question; in other words, identifying the hazard was easiest, identifying its location less so, and predicting what happened next was the most difficult. They argued that the "What Happens Next?" test provided a finer-grained insight into hazard perception ability; for instance, the results revealed that early detection of a

hazard does not necessarily mean accurate later prediction, while the reaction time paradigm primarily measures early detection.

Castro et al. (2014) validated the "What Happens Next?" paradigm with a Spanish driving population, recruiting learners, novices and experienced drivers, as well as a group of recidivist drivers who had multiple traffic offenses and lost all the points on their licenses. They used a similar procedure to Jackson et al. (2009)'s cut to black condition, except that the second, location question involved a three-option multiple choice (the left, right, and middle parts of the scene) rather than free response. While Jackson et al.'s study used only clips where the potential hazard did eventually materialize, Castro et al. also included clips of quasi-hazardous situations, where the potential hazard did not eventually materialize; for these clips, the correct answer for the third, prediction question was that the hazard did not occur. Furthermore, both hazardous and quasi-hazardous situations contained both behavioral prediction (BP; where the precursor and hazard are the same road user) and environmental prediction (EP; where the precursor and hazard are different) situations (Crundall et al. 2012; see Section 1.1.3 for more detail). For hazardous situations, experienced drivers outscored novices, who in turn outscored learners, but there was no difference between recidivist and non-recidivist drivers. All participants also had higher scores for BP than EP clips. For quasi-hazardous situations, experienced drivers outscored learners, and non-recidivists outscored recidivists, but there were no differences between BP and EP scores. While this generally validated the "What Happens Next?" paradigm for a Spanish population, the results were somewhat mixed as non-recidivists might be expected to outscore recidivists in both types of situations. These results also suggested that the type of hazard did not

interact with experience, which seems to refute Crundall et al. (2012) and indeed is contradicted by the study below.

Crundall (2016) conducted three experiments using the "What Happens Next?" test, and found that experienced drivers outperformed novices in all three. The procedure for all experiments was similar to the ones used by Jackson et al. (2009), but clips which did not contain a hazard also made up a quarter of the stimulus pool. In the first experiment, Crundall varied the length of the clips, using short, medium, and long clips. All drivers had the lowest scores on the long clips, although there was no difference between short and medium clips. The drop in performance from medium to long clips was also greater for novices, compared to experienced drivers, possibly suggesting that hazard perception is more effortful for novices. In the second experiment, the cutoff points of the clips were varied, using early (when only the precursor was visible), intermediate (when the precursor was developing into a hazard), and late (immediately prior to hazard onset) cutoff points. As might be expected, clips with early cutoffs were the hardest for all drivers, then intermediate, then late, although this did not interact with experience. In the third and final experiment, BP and EP clips were used, building on Crundall et al. (2012)'s results, which suggested that novice drivers are less likely to fixate BP precursors and EP hazards than experienced drivers. All drivers had higher scores on BP clips, but this was driven by novices' particularly low scores on EP clips, both compared to their scores on BP clips and experienced drivers' scores on EP clips. While this supports Castro et al. (2014) in terms of BP hazards being easier to predict, Castro et al. (2014)'s results also suggested that ease of prediction for BP/EP hazards did not vary with experience.

1.1.4.3 *Static tasks*

In an eye tracking study, Huestegge et al. (2010) recruited novice and experienced drivers to complete a static hazard perception task. Drivers viewed pictures of different road scenes with low, medium, or high braking affordance, where the low affordance scenes had received few braking responses in a previous study, and the medium affordance scenes had received longer reaction times for braking responses than the high affordance scenes. They responded to scenes that they felt required a braking response or speed reduction with a button press as fast as possible. As expected, low braking affordance scenes had significantly less responses than medium affordance scenes, which in turn had less responses than high affordance scenes; all subsequent analyses were conducted with only medium and high affordance scenes. Additional analyses suggested that high affordance scenes received considerably more attention than medium; all drivers responded to high affordance scenes faster, took less time from scene onset to fixate on hazards in high affordance scenes, and pressed the response key faster after hazard fixation. Drivers also had fewer fixations on high affordance scenes than medium, and had longer saccade amplitudes, showing the effect of attentional narrowing found by several other researchers (Chapman & Underwood, 1998; Crundall & Underwood, 1998). However, unlike past findings, attentional narrowing, and in fact all the scene-specific effects, had no interaction with experience. Experience also showed a slightly different pattern of results than scene type. Experienced drivers responded to scenes faster than novices, fixated hazards after scene onset equally quickly, and crucially, pressed the response key faster after hazard fixation, suggesting that their superior reaction time stemmed from processing hazards faster rather than fixating them sooner.

1.1.5 Training

Many studies have reported that hazard perception can be improved through training, although the subsequent improvements in response latency have yet to be linked to later accident rates. Commonly used methods include commentary training (Castro et al., 2016; Crundall, Andrews, van Loon, & Chapman, 2010; Isler, Starkey, & Williamson, 2009; McKenna, Horswill, & Alexander, 2006; Poulsen, Horswill, Wetton, Hill, & Lim, 2010; Wallis & Horswill, 2007; Wetton et al., 2013), the Act and Anticipate Hazard Perception Training program (AAHPT;(Borowsky, Shinar, & Oron-Gilad, 2010; Meir, Borowsky, & Oron-Gilad, 2014), and the risk awareness and training program (Fisher, Pollatsek, & Pradhan, 2006; Pollatsek, Narayanaan, Pradhan, & Fisher, 2006). For a more detailed overview, see McDonald, Goodwin, Pradhan, Romoser, and Williams (2015).

Isler et al. (2009) explored the effects of self-generated commentary training on novice and experienced drivers, where participants were asked to generate a running verbal commentary while watching traffic simulations, identifying any potential and/or immediate hazards they saw. Their performance was measured using a dual hazard perception task, where the primary task was to detect hazards while performing a concurrent secondary tracking task. In the primary task, participants identified hazards by clicking the mouse and immediately saying what the hazard was out loud. In the secondary task, they simulated the steering of a car, also via the mouse, where they had to keep a moving target dot within a small square. Baseline tests conducted before training found that the trained novice group had significantly slower hazard response times and detected less hazards than the trained experienced driver group, but improved in post-training trials, where their performance was roughly equal. Trained novices were also faster and detected more hazards than

untrained novices in post-training trials, despite showing similar performance in the baseline trials. Trained novices also outperformed trained experienced drivers in the secondary task in both pre- and post-training trials, indicating that the commentary training improved only the hazard perception component and not the secondary tracking task.

Wetton et al. (2013) investigated a combination of methods using a mixed training program, with novice drivers undergoing one of four training packages: 1) prediction-based or "what happens next" training, where drivers watched videos that were stopped at a given point (usually immediately before hazard onset), were asked to predict what might happen next, and subsequently received feedback on their responses; 2) expert commentary training, where trainees watched road footage while listening to an expert driver's commentary on the events taking place; 3) hybrid commentary training, where trainees first generated their own commentaries before listening to expert commentaries on the same footage; and 4) full training, which included all three previous interventions. A control group also watched a video containing segments of a learner driver education DVD. To measure the effect of these interventions, three hazard perception tests were administered: one before any training had occurred (*Baseline HPT*), one immediately after training (*Immediate HPT*), and a final HPT approximately one week after training, ranging from 6 to 36 days (*Delayed HPT*). Novice drivers who underwent the full training package significantly improved their hazard perception response times compared to the control group, in both the Immediate and Delayed HPTs. All three individual components ("what happens next", expert commentary, and hybrid commentary) also improved response times in the Immediate HPT but not the Delayed HPT, although some components seemed more beneficial than others. For instance, the addition of the

self-generated commentary did not appear to have any demonstrable benefit, as the expert and hybrid commentary groups showed roughly similar performance, although both outperformed the "what happens next" group. Finally, there was a significant decay in training for all groups from the Immediate to the Delayed HPT, to the point where only the full training condition showed an improvement from baseline in the Delayed HPT.

While many studies have used hazard response time as a primary outcome of training effectiveness, other studies have also found improvements in other areas. Fisher et al. (2006) trained novice drivers using a PC-based risk awareness and training perception program (RAPT), and found that the program improved participants' scanning behavior, as they were more likely to check critical areas that could reduce their likelihood of a crash (see Section 1.1.6 below for a more detailed review). Castro et al. (2016) explored the effects of commentary training on drivers' ability to predict hazards on the road. They recruited learner drivers, novice drivers, and experienced drivers. Performance was measured by performance in the "What Happens Next?" test (described above in Wetton et al. (2013); also see Section 1.1.4.2 above), where drivers watched videos of hazards that were occluded immediately before hazard onset, and answered 3 questions at the end of each clip: (1) What is the hazard? (2) Where is it located? and (3) What happens next? Participants completed a pre-test and post-test version using this paradigm; both versions included gradual onset and abrupt hazards, which used the same definition as Underwood et al. (2013): hazards where the road users were on screen for more than 3 seconds were classified as gradual onset, while less than 3 seconds were considered abrupt hazards. During the training session, participants watched the full, non-occluded pre-test videos with a voice-over describing relevant details of the scene and commenting on potential

dangers on the road, effectively walking participants through the cues that allowed them to predict upcoming hazards. The control group took a break during this time. After another brief break, both groups then completed another version of the "What Happens Next?" test, using the same procedure but with different videos. Castro et al. reported that participants found abrupt hazards easier to detect than gradual onset ones. While both trained and control groups improved scores from pre- to post- test, the trained group outscored controls only on the post-test, suggesting training did indeed have an impact. Training seemed to benefit prediction of both abrupt and gradual hazards and the trained group showed a larger improvement in gradual hazard scores than abrupt, although this may have been a function of the test and/or hazards used, since the control group also significantly improved their prediction of gradual hazards (although not as much as the trained group). Training also benefitted all driver groups, and this effect seemed to grow with driving experience; experienced drivers improved the most, then novices, then learners. Interestingly, however, experienced drivers outperformed learners on only the post-test, which may suggest that this particular test differentiated experience less effectively than past "What Happens Next?" tests (Castro et al., 2014; Jackson et al., 2009; Lim et al., 2014).

Training one aspect of driving also may improve several others. McKenna et al. (2006), reported three experiments where skill-based training appeared to reduce risk-taking behavior. In Experiment 1, novice drivers completed several risk-taking questionnaires and four video-based driving tests: the close following test, where drivers watched a camera car approaching the back of another and indicated following distances at which they were comfortable and uncomfortable; the video speed test, where drivers reported their preferred driving speed relative to the camera car; the gap acceptance test, where drivers watched footage of oncoming traffic and indicated any

gaps they felt acceptably wide to drive into; and lastly, the hazard perception test, where drivers' response time to hazards was recorded. These four tests were conducted after a short video session, where the trained group watched road footage while listening to expert commentary and the control group watched the same footage without commentary. The trained group had shorter response latencies on the hazard perception test and also showed less risk-taking behavior in all relevant measures (following distance, speed, and gap acceptance; some of the risk-taking measures were combined due to high correlations between them). Interestingly, this was not the case for self-perceived skill rating. In Experiment 2 the same training procedure was conducted with another group of novice drivers, but participants completed only the video speed test of the four original post-tests, which had been modified to include scenes with and without explicit hazards. While all drivers responded to training and hazard presence by reporting lower preferred speeds, an interaction was found where trained drivers chose significantly lower speeds in the videos with a hazard present, suggesting that training had improved drivers' ability to detect hazards, rather than lowering their general risk-taking propensity. Finally, Experiment 3 was conducted with police officers who had either passed or not yet taken the advanced UK police driver training course, which involves four weeks of theoretical and practical training, including on-road self-generated commentary training to develop hazard perception skills. Because the purpose of Experiment 3 was to examine whether a non-lab-based training program produced similar results as the commentary training used in Experiments 1 and 2, the training commentary was not used and participants completed only the video speed test used in Experiment 2. They also rated each of the scenarios in a second viewing of the videos. The results mirrored those of Experiment 2's, where the advanced police drivers chose significantly lower speeds in

the videos with a hazard present compared to the non-advanced drivers. Hazard ratings also largely reflected speed choices, where advanced drivers rated the hazard-present videos as more hazardous than the hazard-absent videos. These results suggest that hazard perception training can also decrease risk-taking behavior as well as improve hazard perception itself.

While McKenna et al. (2006)'s results indirectly suggest that training can have long-term effects on hazard perception skill, the advanced police driver training course lasts four weeks, which is impractical for most drivers. In Wetton et al. (2013)'s training program, reviewed above, training effects decayed significantly after one week had elapsed; in fact, performance declined to its pre-training baseline for all groups except those who had received the full training package. However, Horswill, Falconer, et al. (2015) reported that after participating in a brief lab-based training program, older drivers' hazard perception scores showed improvements over baseline 1 and 3 months later, with no significant training decay. Participants were all 65 or older, and were assigned to one of three groups: a non-booster training group, a booster training group who received the same treatment as the first but with an additional booster session one month after initial training, and finally a control, who listened to a driving instructor discussing safe driving that was unrelated to hazard perception. Both initial and booster training involved two different exercises, focused on commentary and predictive methods respectively. Participants completed four hazard perception tests; a baseline HPT administered before any training took place, an HPT conducted immediately post-intervention, a 1-month HPT conducted one month after the initial training session (after which the booster training session was administered where appropriate), and finally, a 3-month HPT conducted three months after the initial training session. Horswill, Falconer, et al. reported that both trained

groups outperformed the control group in all three post-intervention HPTs relative to baseline, showing a clear training effect. They also found that the booster session did not appear to improve performance, as the booster and non-booster training groups showed similar performance relative to baseline in all post-intervention HPTs. Finally, they reported no significant training decay, as performance remained similar relative to baseline across all post-intervention HPTs. The last two results are somewhat unexpected, particularly the lack of training decay, given that Wetton et al. (2013) observed significant decay after just one week. However, as Horswill, Falconer, et al. point out, in a previous study with mid-age drivers, they similarly observed no training decay after one week (Horswill et al., 2013). Given that Wetton et al. (2013)'s participants were mostly university undergraduates obtaining course credit, it seems possible that the mid-age and older drivers may simply have been more motivated. It is also possible that training benefits may simply endure for longer in experienced drivers, because they have a stronger foundation of skill to build upon.

1.1.6 Visual search in novice and experienced drivers

In addition to behavioral studies of hazard perception, a number of studies have utilized eye tracking to understand more about the cognitive mechanisms underlying hazard perception test performance. Chapman and Underwood (1998) showed videos of driving situations to both novice and experienced drivers while recording their eye movements, and asked them to press a response button as soon as they detected a hazard. The clips contained footage from rural, suburban, and urban roads in the UK, with visual complexity generally increasing from rural to urban areas. A "danger window" was defined for each hazard occurring in the films, which began 1 second before the hazard appeared and ended when the road users involved

were no longer visible; participants' response time was measured from the beginning of this window. They reported that during the danger windows, participants' fixation durations increased, particularly among novice drivers, mean saccade angular distances decreased, and both the horizontal and vertical spread of visual search decreased. Overall this suggests a narrowing of attention during the danger window, which Chapman and Underwood observed was similar to that seen in other domains such as eyewitness testimony, where the phenomenon of "weapon focus" occurs (when one's attention focuses on emotionally arousing objects or situations and impairs memory for other details; e.g. (Loftus, Loftus, & Messo, 1987). Attentional focusing seems particularly prominent among novice drivers, who also appear to process visual scenes more slowly given their generally longer fixation durations. Chapman and Underwood also found that for all drivers, fixation duration decreased with road complexity while saccade angular distances increased, suggesting a strategy of rapid, frequent fixations on urban roads and longer fixations on rural. As mentioned in Section 1.1.3, interestingly, there was no difference between novice and experienced drivers' response times to hazards despite the experienced group having held their licenses for 5 – 10 years.

In another study investigating eye movements and road complexity, Crundall and Underwood (1998) asked participants to drive a 20-minute route while wearing a head-mounted eye tracker. They selected three windows for analysis during this drive, each lasting one minute and representing different degrees of complexity; from least to most demanding, a rural single-lane carriageway, a suburban road through a small village, and a dual carriageway with merging traffic. They found that experienced drivers increased their horizontal spread of search on the dual carriageway and to a lesser extent the suburban road, while novices maintained

similar search variance for all three road types. While Chapman and Underwood (1998)'s study above found that novices had generally longer fixation durations, Crundall and Underwood reported different patterns for the driver groups depending on road type; experienced drivers had the longest fixation durations on the rural road and novices on the dual carriageway, although both driver groups showed similar (and the shortest) fixation durations on the suburban route. Crundall and Underwood suggested that novices' longer fixation durations on the dual carriageway might signify the attentional narrowing reported by Chapman and Underwood above, and that experienced drivers had more flexible visual strategies compared to novices.

Falkmer and Gregersen (2005) conducted a similar on-road study, equipping novice and experienced drivers with a head-mounted eye tracker as they drove a 30-minute route through urban, suburban and rural areas. They selected two particular windows for analysis: a city route, where participants drove through city traffic at low speeds, through an area with a zebra crossing, and a rural route, where participants drove on a dual carriageway at higher speeds, eventually passing through a four-way intersection. Falkmer and Gregersen examined the number of fixations made by participants in several areas. They reported that while the vast majority of fixations were on objects further than 5m from the vehicle, experienced drivers made more fixations on closer objects than novices in the city route but not the rural. Novices also made more fixations on in-vehicle objects (all objects closer than 0.6 m for analyses purposes) and the dashboard than experienced drivers did. Mirroring Crundall and Underwood (1998)'s findings of greater flexibility, experienced drivers appeared to adapt their dashboard fixations while novices did not, making more dashboard fixations while driving the rural route than the city. Similarly, experienced drivers also appeared to adapt their fixations according to the route, as they made

relatively more fixations on the roadside immediately beside the car while driving the city route than the rural, while novice drivers did not vary their fixations on the roadside regardless of which route they drove. Novices also fixated the roadside relatively more times than the experienced drivers, which in turn reduced their number of fixations on the road ahead. Finally, novices fixated traffic-relevant objects and potential hazards more than experienced drivers did. It should be noted however that Falkmer and Gregersen analyzed only absolute number of fixations, which might somewhat limit interpretation; for instance, if novice and experienced drivers made the same number of fixations in certain areas, one group may have had fixations of shorter duration and therefore spent less total time fixating an area.

Although novices appear to exhibit some limitations in search strategy compared to experienced drivers, several researchers have reported that scanning behaviors can be improved through training. Fisher et al. (2006) conducted three novice driver experiments using a PC-based risk awareness and training program (RAPT), which trained drivers to recognize areas that posed a particular risk and could provide information that might prevent a potential crash. In the first experiment, trained and untrained novices drove simulated routes while their eye movements were recorded; this drive took place immediately after the RAPT for the trained group. Drivers' eye movements were then analyzed to measure fixations on areas that contained potential threats. Fisher et al. reported that trained drivers fixated on relevant areas significantly more than untrained drivers did. These scanning improvements also appeared to generalize, as trained drivers showed similar improvements in both near transfer scenarios (which were similar to scenarios they had viewed during training), and far transfer scenarios (which were dissimilar). The second experiment used a similar procedure except that the simulated drives took

place 3 – 5 days after training, and produced the same results and a similar effect size to the first experiment. The third experiment was conducted on-road, where eye movements were monitored using a head-mounted eye tracker. Again, trained drivers fixated areas that could reduce crash likelihood significantly more than untrained drivers did, but showed slightly less improvement over untrained drivers for far transfer scenarios compared to near transfer scenarios.

Chapman et al. (2002) also reported changes in novices' scanning behavior several months after a training intervention. The program targeted three skills: knowledge of road situations, scanning behavior, and anticipation on the road. Training exercises included self-generated and expert commentary, "what happens next?" tasks, and highlighting of critical scene areas. Participants were tested at three intervals after passing their driving test: as soon as possible (baseline test phase), three months later (immediate test phase), and 6 – 9 months later (delayed test phase). At each testing phase, all participants completed a hazard perception test and an on-road drive in an instrumented vehicle while wearing a head-mounted eye tracker. The trained group also completed their program at the three month interval. For on-road search, the trained group showed wider horizontal search in general than the control group in the immediate phase, although this effect did not persist to the delayed phase. The video analyses showed within-subject effects consistent with previous studies (Chapman & Underwood, 1998; Crundall & Underwood, 1998): during danger windows, all participants showed longer fixation durations and narrowed horizontal search, which was the case for all three testing phases. The trained group also had shorter mean fixation durations than controls, but in only the immediate phase. They also showed wider horizontal search than the controls in both immediate and delayed phases. This suggests that the intervention had some effect, particularly because the

control group had fairly similar fixation durations and horizontal spread during all three phases, while the trained group showed shorter fixation durations and wider horizontal search. However, as Chapman et al. point out, these scanning behaviors were consistent regardless of the level of danger present, suggesting the trained novices had adopted more conscious scanning behaviors rather than developing the adaptability that often characterizes experienced drivers' search strategies (Crundall & Underwood, 1998; Falkmer & Gregersen, 2005). Training effects also seemed to manifest more strongly when watching videos than on-road driving, suggesting that it may be beneficial to allow a longer period for drivers to develop vehicle control skill, before training higher-order skills. Unfortunately, Chapman et al. did not report video hazard response times, making it difficult to judge whether the scanning improvements accompanied similar improvements in response latency, which would have been especially interesting given Chapman and Underwood (1998) observed visual search differences with no latency differences.

1.2 Road safety in developing countries

1.2.1 Traffic psychology in developing countries

Notably, all the hazard perception research reviewed above has been conducted in developed countries, where road safety is relatively mature. However, the vast majority of road fatalities worldwide occur in developing countries (Peden et al., 2004; Toroyan, 2009, 2013), despite having a lower proportion of registered vehicles. Traffic psychology research in developing countries also largely focuses on the social and practical aspects of driving, rather than the cognitive: for instance, attitudes about seat belt use (Hauswald, 1997), speed choices (Ghadiri, Prasetijo,

Sadullah, Hoseinpour, & Sahranavard, 2013; King, Lewis, & Abdul Hanan, 2011), anger and aggression (Sullman, Stephens, & Yong, 2014, 2015), or evidence to support policy changes (Mohamed et al., 2012; Yusoff et al., 2010), to name a few. Research comparing developed and developing countries is also relatively rare; while some cross-cultural research has been conducted, this has primarily examined risk-taking behavior and/or self-reported perceptions of driving skill and traffic risk (Lund & Rundmo, 2009; Nordfjærn & Rundmo, 2009; Özkan, Lajunen, Chliaoutakis, Parker, & Summala, 2006a, 2006b; Sivak, Soler, Tränkle, & Spagnhol, 1989). Hazard perception has yet to be investigated cross-culturally in depth, even within developed countries, and little is known about the transferability of hazard perception skills between noticeably different countries and cultures; are there crucial underlying skills that can successfully transfer between countries, or are strategies and skills culturally distinct?

Exploring hazard perception cross-culturally can also shed light on how location familiarity impacts hazard perception skill. Wetton et al. (2010) found novice/experienced latency differences in Australian participants when using footage of both Australian and UK roads, suggesting the advantage of experience endures even in unfamiliar environments and hazard perception abilities contain at least some general component. However, several questions remain unanswered. For instance, the UK and Australian settings used by Wetton et al. are very similar. Cultures, road laws, vehicles, driving styles, and even architecture overlap considerably between the two countries. Would similar transference of skill occur in vastly different settings? If a locational advantage does exist, is it due to familiarity with the driving environment itself, familiarity with hazards typically encountered in that environment, or, likely, some combination of the two?

1.2.2 Cross-cultural comparisons: Malaysia and the UK

1.2.2.1 *Road safety statistics*

In this thesis, we will investigate hazard perception skill transferability across cultures: specifically, the UK and Malaysia. As a former British colony, Malaysia's road system shares some similarities with the UK; for instance, the country retains similar road rules and uses left-hand driving. It is also a middle-income country with a high percentage of car ownership; as of 2010, the population of the country was 28.4 million with 20.2 million registered vehicles recorded, or 711 vehicles per 1000 people (Toroyan, 2013). In contrast, in 2010 the UK had a population of 62 million and 35.1 million registered vehicles, or 567 vehicles per 1000 people.

Vehicle composition varies considerably between the two countries, which is roughly reflected in the categories of road user deaths. For instance, in the UK, four-wheeled vehicles account for over 90% of registered vehicles, while powered two- and three-wheelers account for 3.6%. Drivers and passengers of four-wheeled vehicles account for 48% of UK road user deaths, and two- and three-wheelers account for 22%. In Malaysia on the other hand, the proportion of four-wheeled vehicles drops to 45%, and powered two- and three-wheelers make up 47% of vehicle traffic (Toroyan, 2013). Similarly, drivers and passengers of four-wheeled vehicles account for 26% of Malaysian road user deaths, while two- and three-wheelers account for 59%. The greater number of two- and three-wheelers may be one contributor to Malaysia's drastically higher accident rate: Kilbey (2011) reported a road fatality rate of 3 per 100,000 people in the UK in 2010, while Rohayu, Sharifah Allyana, Jamilah, and Wong (2012) reported 24 per 100,000 people in Malaysia. Rohayu et al. (2012) also estimated that total road fatalities would increase from 6,872 in 2010 to 10,716 in 2020. Furthermore, Malaysia does not provide death

registration data, thus it is possible that many fatalities and accidents go underreported. The Malaysian government has established road safety as a priority for the country, holding a national level launch for the UN's Decade of Action of Road Safety 2011 – 2020 program and planning a schedule of activities accordingly. In 2007, the Malaysian Institute of Road Safety Research (MIROS) was set up and more recently, the Road Safety Department announced the Road Safety Plan of Malaysia 2014 – 2020. The goal of the Road Safety Plan is to reduce the projected road fatality rate by 50%, from the original estimate of 10,716 per 100,000 people to 5,358. The plan consists of five strategic pillars: road safety management, safer mobility and roads, safer vehicles, safer road users, and post-crash management.

The current driver licensing process in Malaysia does not require a hazard perception test or any form of video-based testing. Prospective drivers must first pass a written test of road rules to obtain a learner's license. After holding a learner's license for a month, they may then take a practical on-road test, which, when passed, grants a probationary license (effectively, the equivalent of Australia's provisional license; the UK has no such equivalent). After two years, a probationary license may be upgraded to a full, unrestricted license with no further requirements, assuming the holder has not been involved in enough traffic violations to lose all the points on their license.

1.2.2.2 *Other cultural differences*

A quick glance at the pillars and initiatives in the Road Safety Plan reveals some of Malaysia's priorities for road safety: for instance, emergency responses are considered a high priority as one of the plan's five pillars, with several related initiatives such as a first responder community program, and an overall reduction in emergency response time. Other areas highlighted for improvement involve road

infrastructure and vehicle roadworthiness. The stated goals in the "safer road users" pillar also provides some suggestions as to common road user infractions in Malaysia: speeding violations, red light running, helmet use, and seat belt use. Further insight into Malaysia's road safety priorities can be obtained from a glance at the recent MIROS research reports, where a large number of reports concern motorcyclists. Other topics include crash injuries, red light violations, automated enforcement systems, and traffic offenses during country-wide holiday periods such as Chinese New Year and Hari Raya, when a significant proportion of the population travels within the country.

The composition of road users in Malaysia may also contribute to its driving environment in several ways. For instance, given the prevalence of both cars and motorcycles in Malaysia, it is possible that a larger proportion of road users in Malaysia have experience with both driving cars and riding motorcycles, which has been shown to affect hazard perception positively (Underwood et al., 2013). Furthermore, public transport effectiveness is limited in many parts of the country and taxis are relatively inexpensive: these two factors combined with the higher vehicle prevalence in Malaysia may lead to more people having experience riding in cars, which might in turn affect their pedestrian behavior.

One of the recommendations highlighted in the WHO's Global Status Report on Road Safety concerned enforcement, with the report concluding that "enforcement of strong road safety laws is essential for success," and highlighting that enforcement was considered poor in most countries, including developed. This is certainly a concern in Malaysia, where enforcement of laws concerning key road safety risk factors was rated moderately to not very effective. Seat belt wearing rates, an issue previously highlighted in the Road Safety Plan, are an excellent example of both the

importance and challenge of enforcement; for instance, the national seat belt wearing law was expanded to include rear seat occupants in January 2009, but the WHO reported that only 10% of rear seat passengers in Malaysia wore seat belts. Many other areas exist where enforcement is key but may be lacking; for instance, speeding violations, the use of mobile phones, lane changing, and parking, to name a few.

Within the context of this thesis however, we should note that the term culture primarily applies to driving environment, and the terms "Malaysian driver" and "UK driver" simply refer to someone who learned to drive in that country and still drives there: in other words, a driver whose hazard perception skill arose from one of two distinct traffic environments. While an in-depth exploration of the cultural factors involved is beyond the scope of this thesis, we should certainly not discount their importance in addressing road safety.

1.3 Aims and outline of this thesis

While many factors contribute to road safety, given an eightfold difference in fatality rates between Malaysia and the UK we would expect to see at least some difference in hazard perception skill between the populations. By comparing Malaysia and UK drivers' hazard perception abilities in Malaysian and UK road environments, we should obtain further insight into hazard perception transferability across cultures. Wetton et al. (2010)'s findings certainly suggest some amount of transferability, although this was seen in the UK and Australia where accident rates are very similar. In a more hazardous environment, hazard perception skills are arguably even more critical, although it is also possible that in more disparate driving environments, location familiarity may play a bigger role than Wetton et al. (2010) found.

On a practical level, applying the reaction time test in Malaysia should offer some insight as to the feasibility of using a similar test in Malaysian licensing. Malaysia has comparable road safety records, road rules, and income levels to other countries in Southeast Asia (Toroyan, 2009, 2013), so it is possible that results obtained in a Malaysian population may be generalizable to other countries with similar profiles. Therefore, we shall also consider the practical aspects of implementing such a test. For instance, any potential test must of course differentiate between novice and experienced drivers in Malaysia. It must also be easily deployable and suitable for wide-scale testing.

Chapter 2 presents the results of employing the reaction time test with a Malaysian and UK population, using videos filmed in both countries. This was combined with eye tracking results to assess participants' visual strategies. Chapter 3 also employed eye tracking with UK and Malaysian participants; however, a different paradigm was used: Chapter 3 and all subsequent chapters use the "What Happens Next?" test, a predictive task that requires drivers to choose or describe hazards that would have happened next (see Section 1.1.4, Anticipatory skill for more detail). Except for Chapter 4, all chapters use a multiple choice paradigm that is suitable for mass testing. Chapter 4 compares the multiple choice answer format with a free response format similar to the ones used previously by Castro et al. (2014) and Jackson et al. (2009). Chapter 5 explored the necessity of using videos for correct predictions, and Chapter 6 investigated the effect of altering the amount of visual information present in the videos. Finally, an item analysis was conducted in Chapter 7 after combining the results of Chapters 3, 5, and 6, since these three experiments used the same set of videos and the same task. All experiments were conducted with

novice and experienced drivers. Chapters 2, 3, and 5 used UK and Malaysian participants, while Chapters 4 and 6 used only Malaysian participants.

CHAPTER 2

PILOTING THE REACTION TIME TASK IN MALAYSIA

Adapted from: Lim, P. C., Sheppard, E., & Crundall, D. (2013). Cross-cultural effects on drivers' hazard perception. Transportation Research Part F: Traffic Psychology and Behaviour, 21, 194–206.

Abstract

Hazard perception tests are used in several developed countries as part of the driver licensing curriculum, however little research has been done in developing countries where road safety is a primary concern. We conducted a cross-cultural hazard perception study to examine the transferability of hazard perception skills between Malaysia and the UK, using hazard clips filmed in both countries. The results showed that familiarity with both the driving environment and type of hazard facilitated drivers' ability to discriminate hazards in a timely manner, although overall drivers viewed and responded to hazards largely similarly regardless of origin. Visual strategies also appeared to be moderated mainly by the immediate driving environment rather than driver origin. Finally, Malaysian drivers required a higher threshold of danger than UK drivers before they would identify a situation as hazardous, possibly reflecting the more hazardous road environment in Malaysia. We suggest that hazard perception testing, particularly in developing countries, would benefit from a paradigm where performance cannot be confounded with differing thresholds for hazardousness.

2.1 Introduction

The hazard perception skill of a driver refers to the ability to identify potentially dangerous situations on the road. It is typically assessed by showing participants video clips of hazards, and asking them to respond as soon as they detect a developing hazard, with shorter response times reflecting greater levels of HP skill (Chapman & Underwood, 1998; Horswill & McKenna, 2004; McKenna et al., 2006; Sagberg & Bjørnskau, 2006; Wetton et al., 2011). According to Horswill and McKenna (2004), hazard perception is one of the only components of driving skill that has been consistently linked to accident involvement across multiple studies (Horswill et al., 2010; McKenna & Horswill, 1999; Quimby et al., 1986). A prospective study by Drummond (2000) also found that newly licensed drivers' hazard perception performance was linked to their likelihood of being involved in a fatal collision in the subsequent 12 months. Additionally, past research has found that experienced drivers outperform novices in hazard perception tests (Horswill et al., 2008; Scialfa et al., 2011; Wallis & Horswill, 2007), with driving experience considered as a proxy for driver safety, although other studies have failed to find this experiential difference (Chapman & Underwood, 1998; Crundall, Underwood, & Chapman, 2002; Sagberg & Bjørnskau, 2006). It has recently been suggested that the lack of replication may stem in part from the particular hazards used in these studies, with certain types of hazard differentiating experience more effectively than others (Borowsky, Shinar, & Oron-Gilad, 2007; Crundall et al., 2012). Nevertheless, the reported predictive utility of hazard perception testing has led to the inclusion of a hazard perception component in driver licensing in the UK, Australia and the Netherlands, where there has been some evidence to suggest its efficacy in reducing

accidents in new drivers (Boufous et al., 2011; Horswill, Hill, et al., 2015; Wells et al., 2008).

Notably, the research cited has been conducted in developed countries where road safety is relatively mature, yet the vast majority of road fatalities worldwide occur in developing countries (Nantulya & Reich, 2002; Peden et al., 2004; Toroyan, 2009). While some cross-cultural research on driving has been conducted comparing developed and developing countries (Lund & Rundmo, 2009; Nordfjærn & Rundmo, 2009; Özkan et al., 2006b; Sivak et al., 1989), this has primarily examined risk-taking behavior and/or self-reported perceptions of driving skill and traffic risk. Hazard perception has yet to be investigated cross-culturally in depth, even within developed countries, and little is known about the transferability of hazard perception skills between noticeably different countries and cultures; are there crucial underlying skills that can successfully transfer between countries, or are strategies and skills culturally distinct?

Exploring hazard perception cross-culturally can also shed light on how location familiarity impacts on hazard perception skill. Wetton et al. (2010) found novice/experienced latency differences in Australian participants when using footage of both Australian and UK roads, suggesting the advantage of experience endures even in unfamiliar environments and hazard perception abilities contain at least some general component. However, several questions remain unanswered. For instance, the UK and Australian settings used by Wetton et al. are very similar. Cultures, road laws, vehicles, driving styles, and even architecture overlap considerably between the two countries. Would similar transference of skill occur in vastly different settings? If a locational advantage does exist, is it due to familiarity with the driving environment itself (which we will term *environmental* familiarity), familiarity with

hazards typically encountered in that environment (*hazard* familiarity) or, likely, some combination of the two?

The current study aims to investigate the extent to which hazard perception skills transfer cross-culturally, using the typical reaction time paradigm, across two highly different settings: the UK and Malaysia. This will reveal similarities and differences in hazard perception performance between drivers from two very different driving cultures, and will hopefully identify whether these are specific to the context (i.e. can UK drivers perform well on both UK and Malaysian HP clips?).

Hazard perception performance of experienced and novice drivers from both the UK and Malaysia will be compared using the reaction time paradigm described earlier with video footage from both the UK and Malaysia. As a former British colony, Malaysia shares several commonalities with the UK that make it a suitable comparison point; namely, similar road rules and left-hand driving environment. It is also a middle-income country with a high percentage of car ownership. However, Malaysia has drastically higher accident rates; in 2010 its road fatality rate was 24 per 100,000 people (Rohayu et al., 2012), while the UK's was 3 per 100,000 (Kilbey, 2011), suggesting many more dangerous events and a generally more hazardous road environment. While many factors likely contribute to this discrepancy, given an eightfold difference in fatality rates we expect to see at least some difference in hazard perception skills between the populations. By comparing Malaysia and UK drivers' hazard perception abilities in Malaysian and UK road environments, we should obtain further insight into hazard perception transferability across cultures. Wetton et al. (2010)'s findings certainly suggest some amount of transferability, although this was seen in the UK and Australia where accident rates are very similar. In a more hazardous environment, hazard perception skills are arguably even more

critical, although it is also possible that in more disparate driving environments, location familiarity may play a bigger role than Wetton et al. found.

While we expect to see the typical effect of experience in both locations, similar to Wetton et al. (2010), we also expect location familiarity to confer a significant advantage in these two distinct cultures, and we therefore hypothesize superior performance in the form of shorter reaction times and higher response rates when participants view clips from their home country. Furthermore, while Malaysian drivers certainly experience more hazards than UK drivers, this effect could play out in either direction: they may be quicker to detect hazards due to their greater exposure to them, or equally, they may be desensitized to hazards and have a higher criterion for hazard identification compared to UK drivers.

Additionally, we can break down pure location familiarity and infer its subcategories of environmental and hazard familiarity, as mentioned earlier. Half of the clips from both the UK and Malaysia were matched for hazard content. For example, in one matched pair of clips, a car on the highway overtakes the camera car on the inside lane. The other half of the clips consisted of unmatched hazards that are more representative of the different countries. For instance, motorcycle and scooter riders are relatively infrequent hazards in the UK, while in Malaysia powered two-wheelers make up a significantly greater proportion of the traffic. Similarly, zebra crossings are relatively common in the UK, but extremely rare in Malaysia compared to pelican crossings. We expect both environmental and hazard familiarity to confer an advantage, and hypothesize that all drivers will exhibit superior performance when viewing matched hazards in their home country, compared to matched hazards in their non-home country (environmental familiarity), and also when viewing matched non-home country hazards compared to unmatched non-home country hazards (hazard

familiarity). Furthermore, we hypothesize the greatest performance difference between Malaysian and UK drivers when viewing unmatched hazards, as these presumably confer both environmental and hazard familiarity.

Finally, the current study also investigates visual strategies by using eye tracking measures for all participants. This provides a greater insight into hazard perception skill than a measure of response time can provide. For instance, we expect experienced drivers in particular to have shorter fixation durations in their home environment, indicating greater processing efficiency (Crundall & Underwood, 1998), although it is possible this might be offset by the more cluttered road environment in Malaysia, as more complex environments may also necessitate shortened fixation durations. Experienced drivers should also show wider search patterns along the horizontal meridian, similar to Chapman and Underwood (1998)'s finding. We expect to see the typical effects of attentional capture across all clips, i.e. longer fixation durations upon hazard onset, although this should be less pronounced in experienced drivers (Chapman & Underwood, 1998). It may also help establish whether differences in hazard perception performance stem from differences underlying attentional processes (how quickly participants spot the hazard) or differences in criterion (how quickly participants judge the hazard to pose a threat).

2.2 Methods

2.2.1 Participants

Forty-five participants were recruited from the UK and 55 from Malaysia, all of whom held either full or learner driving licenses[1] from their respective countries

[1] The licensing process in Malaysia and the UK uses slightly different terminology (see Sections 1.1.1 and 1.2.2). A learner's license in Malaysia is the equivalent of a provisional license in the UK, and upon passing the on-road test, UK drivers receive a full license while Malaysian drivers receive a probationary license, which may be upgraded to a full license after two years with no further

and had normal or corrected-to-normal vision. Participants were split into two further sub-groups consisting of novice and experienced drivers, resulting in four groups in total: 20 UK novice drivers (mean age of 18.9 years and licensing time of 9.1 months), 25 UK experienced drivers (mean age of 21.0 years and licensing time of 45.6 months), 26 Malaysian novice drivers (mean age of 18.8 years and licensing time of 9.8 months) and 27 Malaysian experienced drivers (mean age of 21.9 years and licensing time of 49.4 months). Participants received either monetary compensation or course credit, where the latter was applicable.

2.2.2 Stimuli

A Panasonic HD SDC-600 and SDC-900 camera was attached to the windscreen of various cars during journeys made around the UK and Malaysia respectively. Footage was shot in the daytime, under clear weather conditions and normal visibility. Twenty clips from Malaysia and 20 clips from the UK, each containing one hazard and ranging from 6 – 54 seconds in length, were selected and edited from the resulting footage. **Table 2-1** contains clip length information for all clip categories.

Of the 20 clips from each country, 10 were matched for hazard content and what happened in the clip. For example, in one matched pair of clips, a pedestrian steps out from behind a parked bus directly in front of the camera car. The other 10 clips from each country were not matched and reflected hazardous situations that were selected without restriction, and were subsequently more representative of the different countries. For instance, in an unmatched Malaysia clip an oncoming motorcyclist pulls out from behind a truck and into the path of the camera car, while

---

requirements. For consistency, in this thesis we will refer to all drivers who have not passed the on-road test as learner drivers, and licensing time will always be given as the time since passing the on-road test.

in an unmatched UK clip a pedestrian with a bicycle steps into the road at a zebra crossing, forcing the camera car to stop and allow the pedestrian to pass. Matched clip pairs and unmatched clips were selected independently by one Malaysian researcher and one UK researcher, both of whom held a full driving license in their respective countries. Disagreements were resolved and the final clips, both matched and unmatched, selected via discussion between the above two researchers and a further two who had spent a significant amount of time driving in both countries.

Table 2-1: Mean length of clips and hazard windows for all clip categories.

	Malaysian clips		**UK clips**	
	Matched	**Unmatched**	**Matched**	**Unmatched**
Hazard window length (s)	5.19	4.80	4.80	4.90
Clip length (s)	21.4	27.8	34.9	21.9

2.2.3 Apparatus

The stimuli were played on a 17" TFT monitor at a resolution of 1024 x 768, presented using Tobii Studio 2.3. Participants were seated 65 cm from the screen at a visual angle of 29.6° x 23.2 °, and their eye movements were monitored using a Tobii T60 eye tracker in Malaysia and Tobii T1750 eye tracker in the UK, which sample at 60 Hz and 50 Hz respectively. While this results in a slightly higher margin of error for UK participants (20 ms compared to 16.67 ms), this should not introduce any systematic bias into the results.

2.2.4 Design

A 2 x 2 x 2 x 2 mixed design was used. The between-groups factors were the country of origin of the driver (Malaysia or UK), and experience level (novice or experienced). The within-groups factors were the country where the clip was filmed (Malaysia or UK) and whether the clip was matched or unmatched.

The stimuli were separated by country into two blocks of 20 clips, i.e. one Malaysia block and one UK block. Within each block, the order of clips to be presented was randomized using a Latin square. The order of the blocks was counterbalanced across participants.

A *hazard window* was defined for each hazard, similar to the "danger window" used by Chapman and Underwood (1998). A single researcher defined all forty windows, of which a sub-sample were rated by a second researcher with a high degree of agreement. **Table 2-1** above depicts the mean length of each hazard window for each clip category. The window began at the earliest point in time where the hazard was detectable by the viewer and clearly on a course that would eventually obstruct the camera car (the hazard's *onset*). For instance, a pedestrian walking beside the road would not be considered a hazard until he steps towards the road, or makes any other movement that sets him on a collision course with the camera car. The window ended at the point at which a braking or avoidance response by the driver would no longer prevent a collision (the hazard's *offset*). **Figure 2-1** illustrates an example with a Malaysian video.

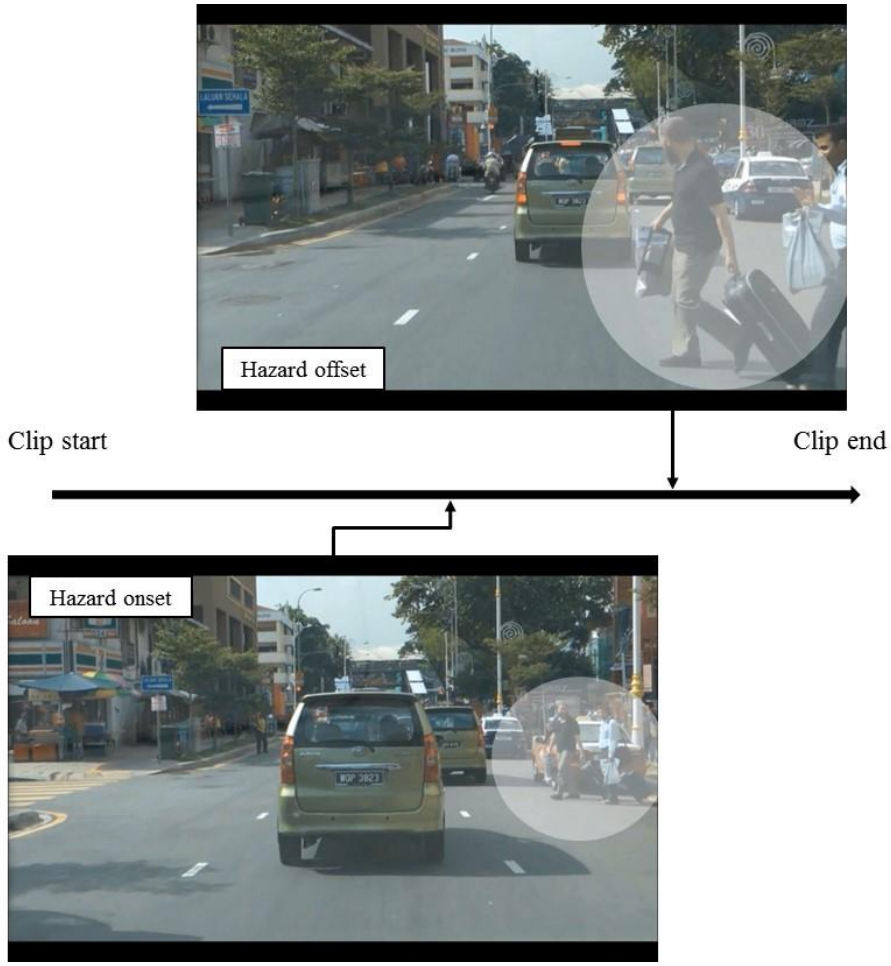

Figure 2-1: Video stills from a sample Malaysian clip illustrating hazard onset, offset, and window. At the hazard onset, the pedestrians pictured have just started to walk across the road in front of the camera car. At the hazard offset, the pedestrians are in front of the camera car, where a braking or avoidance response would no longer avoid a collision. The time elapsed between the offset and onset is defined as the hazard window.

2.2.5 Procedure

After giving informed consent, participants were seated in front of the eye tracker and instructed to watch the video clips and click the left mouse button as soon as they saw a hazard developing. Participants were told to respond to "hazardous events, i.e. situations in which the driver may need to change speed or direction of their vehicle to prevent a potential collision," similar to the definition utilized by Wetton et al. (2011), although the term "hazard" was used instead of "traffic conflict." They were informed that at the end of each video, they would be asked to briefly

describe the last event they had responded to, and then rate its hazardousness on a 1 – 5 scale, where a higher rating indicated a more hazardous event. Participants were then calibrated on the eye tracker, using a standard 9-point calibration procedure. Before starting each 20-clip block, participants were shown two sample clips for them to practice the task, both taken from the same country as the block they were about to view. These practice clips were not used in the subsequent experiment and were the same for all participants. After the practice trials were completed, participants began the first block of clips. Participants were re-calibrated on the eye tracker in-between blocks.

2.3 Results

Four behavioral measures were analyzed: responses to pre-defined hazards (responses were made within the hazard window and the participant verbally reported the appropriate hazard at the end of the clip), number of extra-hazard responses (button presses that were not in response to the pre-defined hazard) for every pre-defined hazard response, reaction time, and ratings of hazardousness. Three eye tracking measures were analyzed: time for participants to first fixate the hazard (regardless of whether they identified it as one), the mean fixation duration prior to and directly after hazard onset, and spread of search along the horizontal meridian prior to hazard onset. Hazard precursors were not included in the analyses.

Unless otherwise stated, a 2 (driver origin: Malaysia or UK) x 2 (experience level: novice or experienced) x 2 (clip country: Malaysia or UK) x 2 (matching: matched or unmatched) mixed ANOVA was run for all measures.

2.3.1 Behavioral analysis

2.3.1.1 *Pre-defined hazard responses*

An analysis of responses to pre-defined hazards (where a correct answer is indicated by a button response during the hazard window and a correct verbal identification following the clip) was conducted. Data for all drivers is summarized in **Figure 2-2**.

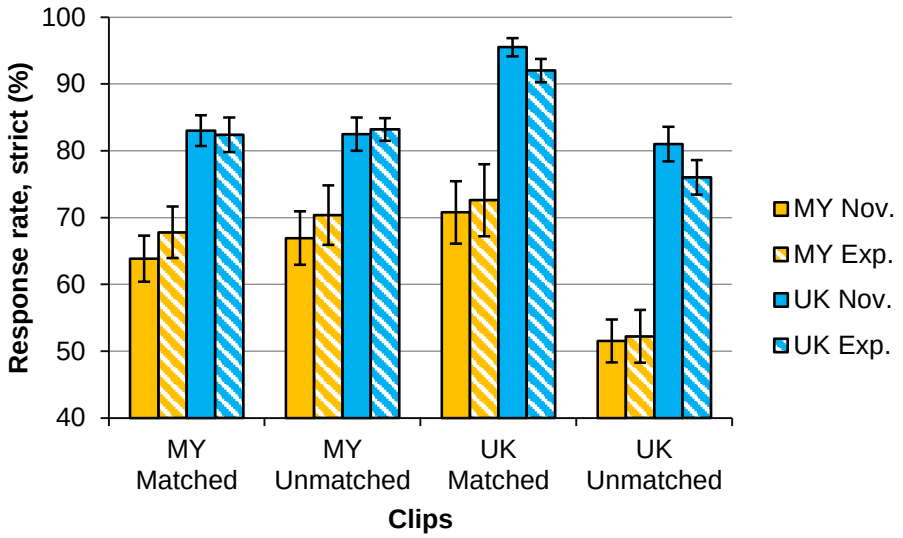

Figure 2-2: Summary data. Response rate based on verbal identification of pre-defined hazards and a button response made during the hazard window. Error bars represent standard error of the mean.

There were two main effects. First, UK drivers identified more hazards than Malaysian drivers ($F_{1,94} = 47.04$, $p < .001$, $\eta^2_\rho = .334$) and secondly, all drivers identified more matched hazards than unmatched ($F_{1, 94} = 64.87$, $p < .001$, $\eta^2_\rho = .408$). There were two two-way interactions, depicted in **Figure 2-3**. Panel (a) shows an interaction between matching and clip country ($F_{1, 94} = 58.43$, $p < .001$, $\eta^2_\rho = .383$), where participants responded equally often to both matched and unmatched Malaysian hazards, but more often to UK matched hazards than UK unmatched ($t(97) = 10.93$, $p < .001$, $d = 0.82$). A second interaction, shown in panel (b), was found

between driver origin and clip country ($F_{1,94}$ = 12.39, p = .001, η^2_p = .116), where participants responded to hazards filmed in their home country more often. This effect was significant in participants from both countries ($t(52)$ = 2.81, p = .007, d = 0.28 for Malaysian drivers and $t(44)$ = 2.19, p = .034, d = 0.40 for UK drivers).

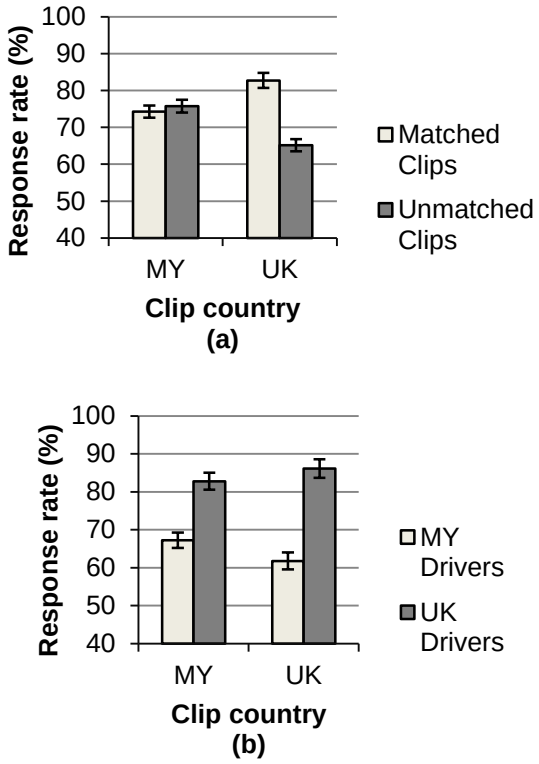

Figure 2-3: Response rate based on verbal identification of pre-defined hazards and a button response made during the hazard window, showing two 2x2 interactions. Panel (a) shows an interaction of matching and clip country and panel (b) shows an interaction of driver origin and clip country. Error bars represent standard error of the mean.

Previous studies have however raised problems with using a restrictive window for calculating responses to hazards, as some exceptional individuals may be penalized for responding too soon, while some drivers may recognize the hazard but forget to respond (Jackson et al., 2009). A second analysis was therefore undertaken which included all trials where participants correctly identified the pre-defined hazard in their later verbal description, regardless of when or whether they had made a button press response during the clip. This, more liberal, analysis was chosen to ensure that

response time outliers did not confound the response rate measure. Data for all drivers is summarized in **Figure 2-4**.

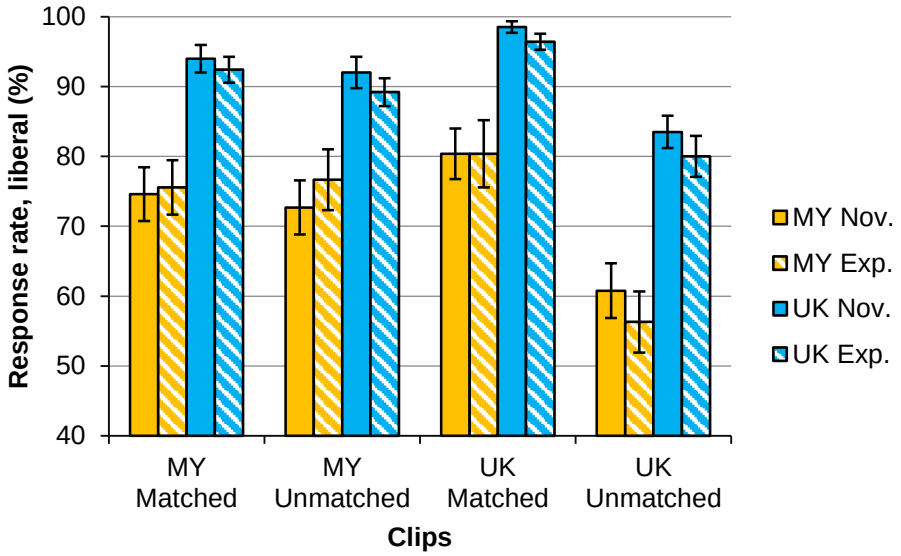

Figure 2-4: Summary data. Response rate based on verbal identification only. Error bars represent standard error of the mean.

As in the previous analysis, main effects of driver origin ($F_{1,94} = 39.33$, $p < .001$, $\eta^2_\rho = .295$) and matching ($F_{1,94} = 112.47$, $p < .001$, $\eta^2_\rho = .545$) were found, where UK drivers identified more hazards, and there were more responses to matched hazards. There was also a main effect of clip country ($F_{1,94} = 13.31$, $p < .001$, $\eta^2_\rho = .124$), where drivers identified Malaysian hazards more often. An interaction between matching and clip country was again found ($F_{1,94} = 71.51$, $p < .001$, $\eta^2_\rho = .432$), showing the same pattern as this interaction in the previous analysis. This was subsumed by a three-way interaction between matching, clip country and driver origin ($F_{1,94} = 4.17$, $p = 0.44$, $\eta^2_\rho = .042$), shown in **Figure 2-5**, where all drivers responded to UK unmatched clips the least compared to the other three clip groups. While this was significant for both groups of drivers when compared to performance on unmatched clips filmed in Malaysia ($t(52) = 6.41$, $p < .001$, $d = 0.76$ for Malaysia

drivers and $t(44) = 4.41$, $p < .001$, $d = 0.77$ for UK drivers), the drop in response rate is numerically greatest in Malaysian drivers.

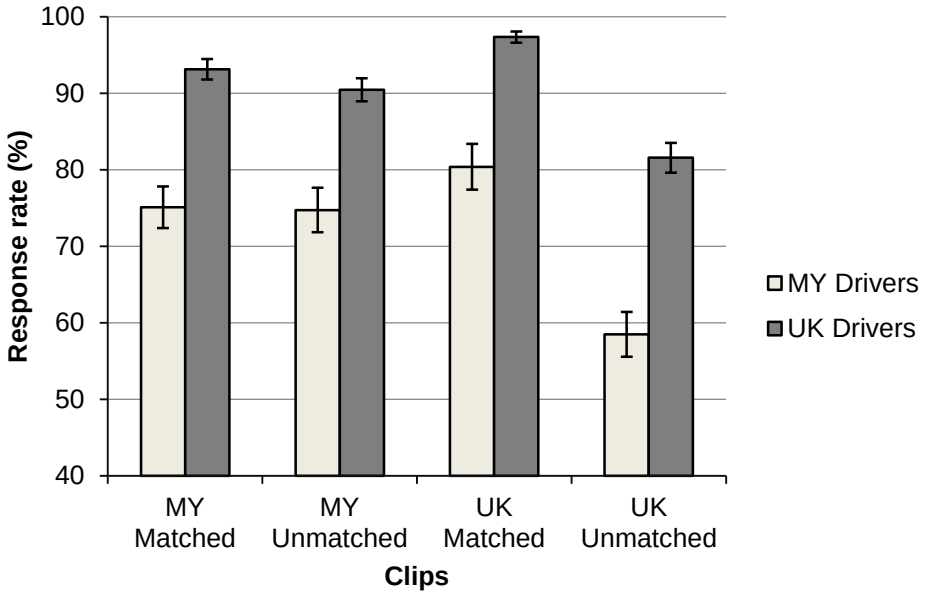

Figure 2-5: Response rate for verbal identification only, showing a 3-way interaction of matching, clip country, and driver origin. Error bars represent standard error of the mean.

2.3.1.2 *Extra-hazard responses to pre-defined hazard responses*

The extra-hazard response rate was the number of button press responses made during an entire video that were not a response to the pre-defined hazard. The number of extra-hazard responses made for every pre-defined hazard response, using the initial strict measure of response rate, was calculated for each participant. A higher number therefore means more extra-hazard responses made relative to pre-defined hazard responses. Data for all drivers is summarized in **Figure 2-6**.

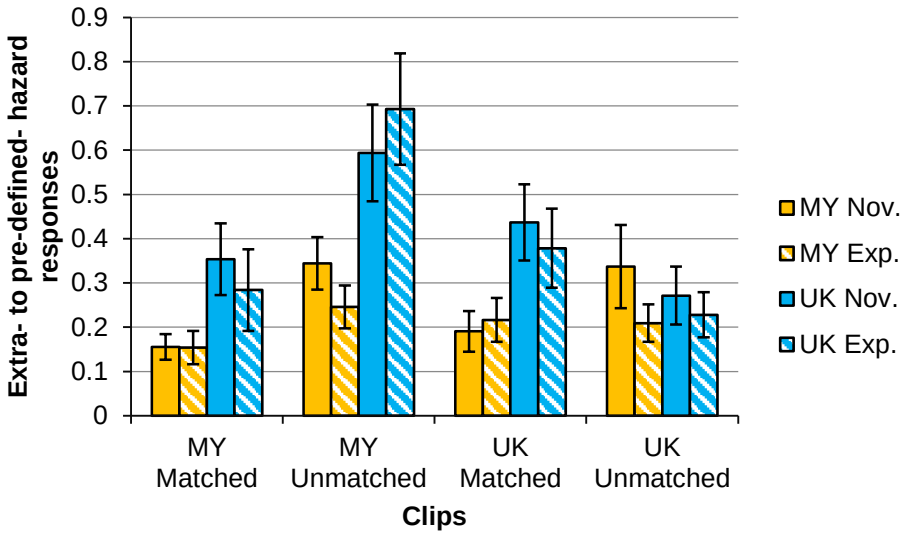

Figure 2-6: Summary data. Number of extra-hazard responses made for every pre-defined hazard response. Error bars represent standard error of the mean.

Three main effects were found: matching, where unmatched clips had a higher rate of responses ($F_{1,94} = 25.15$, $p < .001$, $\eta^2_\rho = .211$), clip country, where Malaysia clips had a higher rate of responses ($F_{1,94} = 6.05$, $p = .016$, $\eta^2_\rho = .060$) and driver origin, where UK drivers had a higher rate of responses ($F_{1,94} = 8.76$, $p = .004$, $\eta^2_\rho = .085$). Two two-way interactions were found, clip country and driver origin ($F_{1,94} = 8.61$, $p = .004$, $\eta^2_\rho = .084$) and matching and clip country ($F_{1,94} = 27.19$, $p < .001$, $\eta^2_\rho = .224$). Both these interactions were subsumed by a three-way interaction of clip country, matching and driver origin ($F_{1,94} = 15.05$, $p < .001$, $\eta^2_\rho = .138$), shown in **Figure 2-7**. For matched clips, all drivers made more extra-hazard responses in UK clips and UK drivers made more extra-hazard responses than Malaysian drivers; however, for unmatched clips, UK drivers watching Malaysian clips had a particularly high extra-hazard response rate ($t(44) = 5.93$, $p < .001$, $d = 0.90$ compared to UK unmatched clips), but there was no difference between Malaysian drivers watching UK and Malaysian clips.

Figure 2-8 shows an additional three-way interaction of driver origin, driver experience level, and matching ($F_{1,94} = 8.37$, $p = .005$, $\eta^2_p = .082$), where novice Malaysian and experienced UK drivers had a higher response rate on unmatched clips compared to unmatched ($t(25) = 4.23$, $p < .001$, $d = 0.72$ and $t(24) = 6.34$, $p < .001$, $d = 0.34$ respectively), but experienced Malaysian drivers and UK novices showed no such difference.

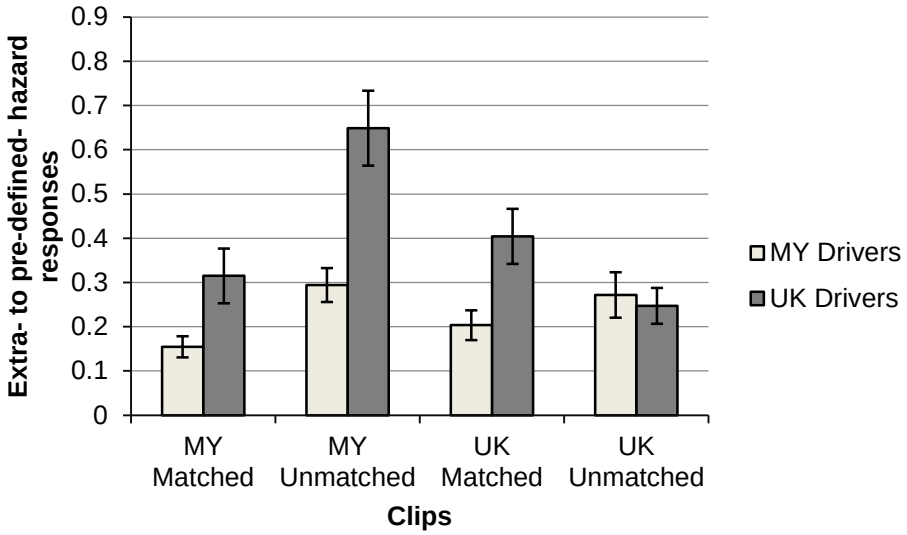

Figure 2-7: Number of extra-hazard responses made for every pre-defined hazard response, showing a 3-way interaction of matching, clip country, and driver origin. Error bars represent standard error of the mean.

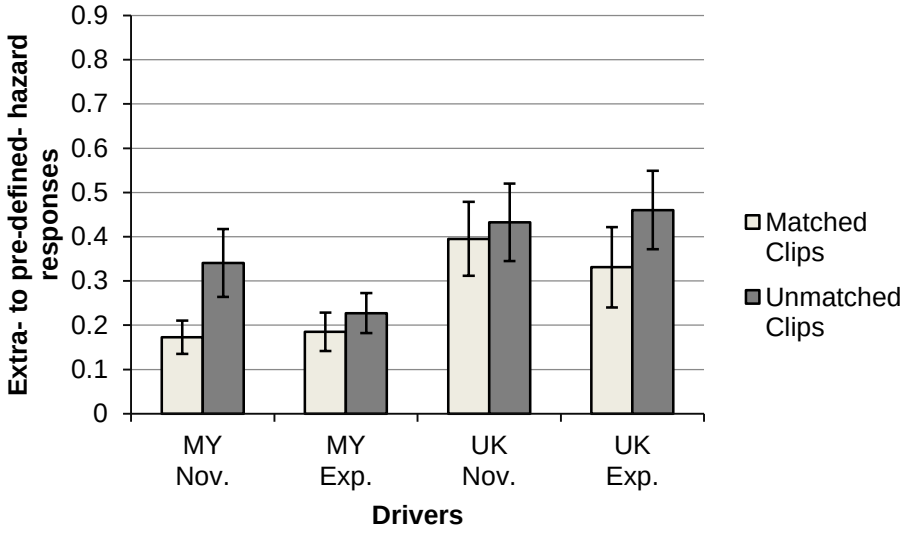

Figure 2-8: Number of extra-hazard responses made for every pre-defined hazard response, showing a 3-way interaction of driver origin, experience level, and matching. Error bars represent standard error of the mean.

2.3.1.3 *Reaction time*

Reaction times were calculated from the beginning of the hazard window, based on the initial strict measure of response rate. Only clips that had responses to the pre-defined hazards during the hazard window were included; all other responses were removed, as were clips with no responses. Data for all drivers is summarized in **Figure 2-9**.

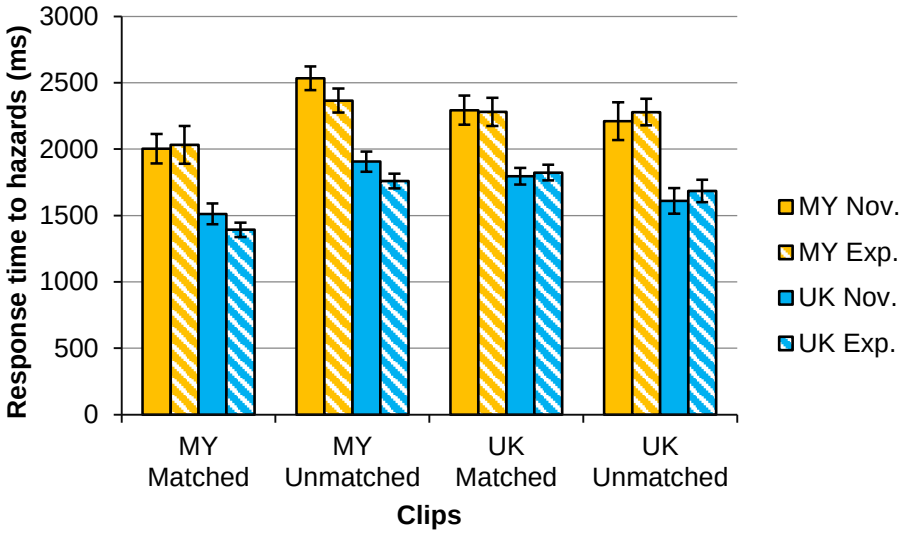

Figure 2-9: Summary data. Reaction time to hazards. Error bars represent standard error of the mean.

Two main effects were found: matching ($F_{1,94} = 20.44$, $p < .001$, $\eta^2_\rho = .179$), where drivers responded to matched hazards faster, and driver origin ($F_{1,94} = 53.94$, $p < .001$, $\eta^2_\rho = .365$), with UK drivers having faster overall response times. **Figure 2-10** shows a crossover interaction of clip country and matching ($F_{1,94} = 49.63$, $p < .001$, $\eta^2_\rho = .346$), where drivers reacted fastest to Malaysian matched clips out of all clips ($t(97) = 5.98$, $p < .001$, $d = 0.57$ compared to UK matched clips) and slowest to

Malaysia unmatched clips ($t(97) = 3.67$, $p < .001$, $d = 1.13$ compared to UK unmatched clips) but there was no difference in reaction time for UK clips regardless of matching.

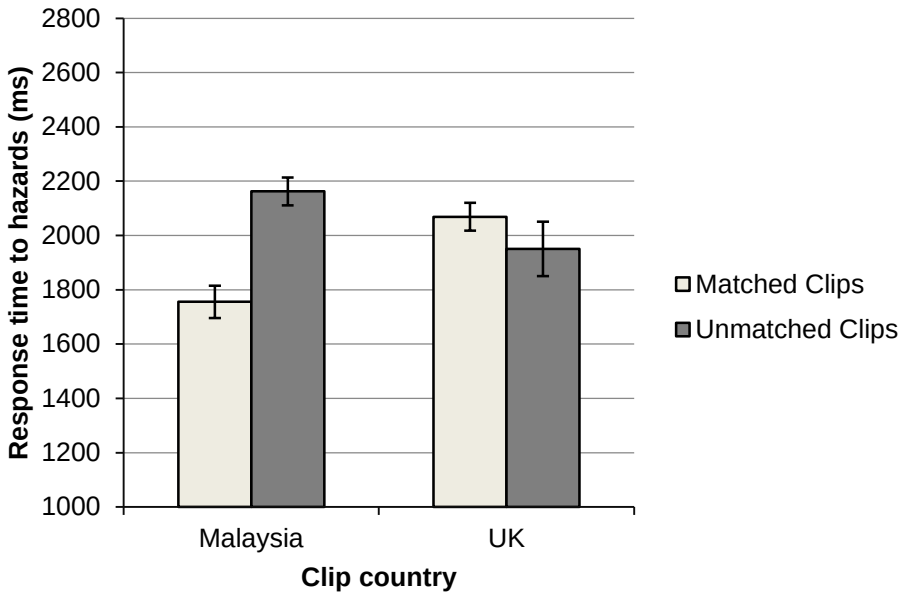

Figure 2-10: Reaction time to hazards, showing an interaction of clip country and matching. Error bars represent standard error of the mean.

2.3.1.4 *Hazard ratings*

Hazard ratings were included in all cases where the participants had correctly identified the hazard, regardless of when or whether they made a button press response. Ratings were removed if participants had rated a different event from the hazard, or had not identified a hazard at all. A higher rating indicates a more hazardous event. Data for all drivers is summarized in **Figure 2-11**.

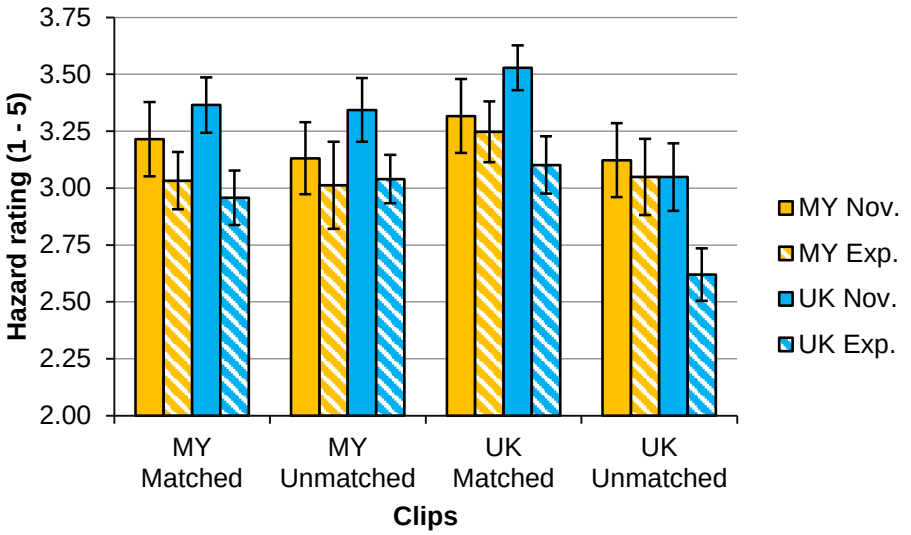

Figure 2-11: Summary data. Ratings of hazardousness. Ratings were from 1-5, with a higher rating indicating a more hazardous event. Error bars represent standard error of the mean.

There was a main effect of matching ($F_{1,94} = 25.5$, $p < .001$, $\eta^2_\rho = .214$), where matched clips were rated as more hazardous. Two crossover interactions were found: one between clip country and matching ($F_{1,94} = 25.9$, $p < .001$, $\eta^2_\rho = .216$) where all Malaysian clips were rated equally regardless of matching, but UK matched clips were rated as most hazardous of all clips ($t(97) = -3.12$, $p = .002$ compared to Malaysian matched clips) and UK unmatched clips least hazardous ($t(97) = 2.59$, $p = .011$ compared to Malaysian unmatched clips). The second interaction was between clip country and driver country ($F_{1,94} = 4.3$, $p = .041$, $\eta^2_\rho = .044$), where drivers rated hazards from their home country as less hazardous, although post hoc t-tests revealed no significant differences. Both these interactions were subsumed by a three-way interaction of matching, clip country and driver country ($F_{1,94} = 8.1$, $p = .005$, $\eta^2_\rho = .080$), shown below in **Figure 2-12**. Both Malaysian and UK drivers rated UK matched clips as more hazardous than UK unmatched clips ($t(52) = 2.86$, $p = .006$, $d = 0.25$ for Malaysian drivers and $t(44) = 7.87$, $p < .001$, $d = 0.78$ for UK drivers) and Malaysian matched clips ($t(52) = 2.28$, $p = .027$, $d = 0.21$ for Malaysian drivers and

t(44) = 2.106, p = .041, d = 0.26 for UK drivers) and found matched and unmatched Malaysian clips equally hazardous. However, UK drivers considered UK unmatched clips particularly non-hazardous (t(44) = 4.92, p < .001, d = 0.59 compared to Malaysian unmatched clips), while Malaysian drivers found unmatched UK and Malaysian clips equally hazardous.

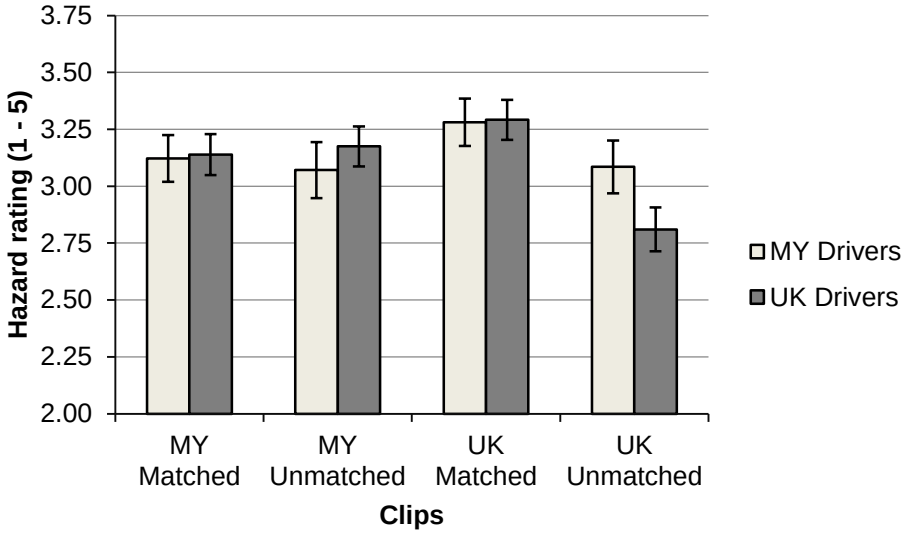

Figure 2-12: Ratings of hazardousness, showing a 3-way interaction of driver origin, experience level, and matching. Error bars represent standard error of the mean.

2.3.2 Eye movement analysis

Eye movement analyses were conducted using Tobii Studio's Dynamic Area of Interest (AOI) tool. To ensure missing eye tracking data did not adversely affect the analysis, only participants who had sampling rates over 65% in at least 90% of the video clips were included. 30 participants were removed according to this criterion, leaving 68 participants in total (16 UK novices; 19 UK experienced; 13 Malaysian novices; 20 Malaysian experienced). All behavioral analyses were re-run with these participants to confirm that the results followed the same pattern as above.

2.3.2.1 *Time to first fixate pre-defined hazards*

The time taken to first fixate on pre-defined hazards was measured from hazard onset, until the participant fixated on the hazard for the first time. If there were no fixations on the hazard during the hazard window, that particular trial was removed. Data for all drivers is summarized in **Figure 2-13**.

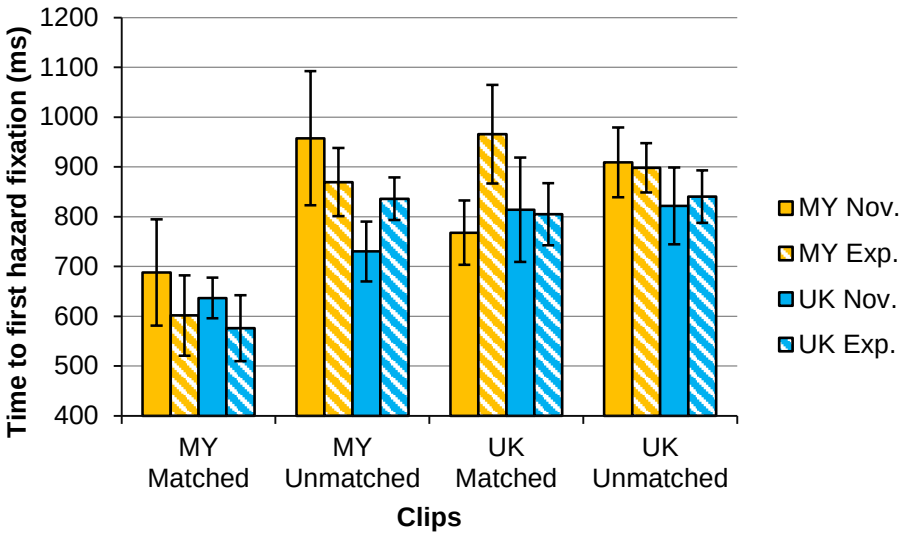

Figure 2-13: Summary data. Time to first fixate hazards. Error bars represent standard error of the mean.

Two main effects and a subsequent interaction were found. Matched hazards ($F_{1,64} = 18.21$, $p < .001$, $\eta^2_\rho = .221$) were fixated faster, as were Malaysian hazards ($F_{1,64} = 10.31$, $p = .002$, $\eta^2_\rho = .139$). This was driven by an interaction of matching and clip country ($F_{1,64} = 11.11$, $p = .001$, $\eta^2_\rho = .148$), depicted in **Figure 2-14**, where Malaysian matched hazards were fixated particularly quickly compared to both Malaysian unmatched hazards ($t(67) = 6.68$, $p < .001$, $d = 0.73$) and UK matched hazards ($t(67) = 4.46$, $p < .001$, $d = 0.68$), but there was no difference in time to first fixate the hazards for all other clip types.

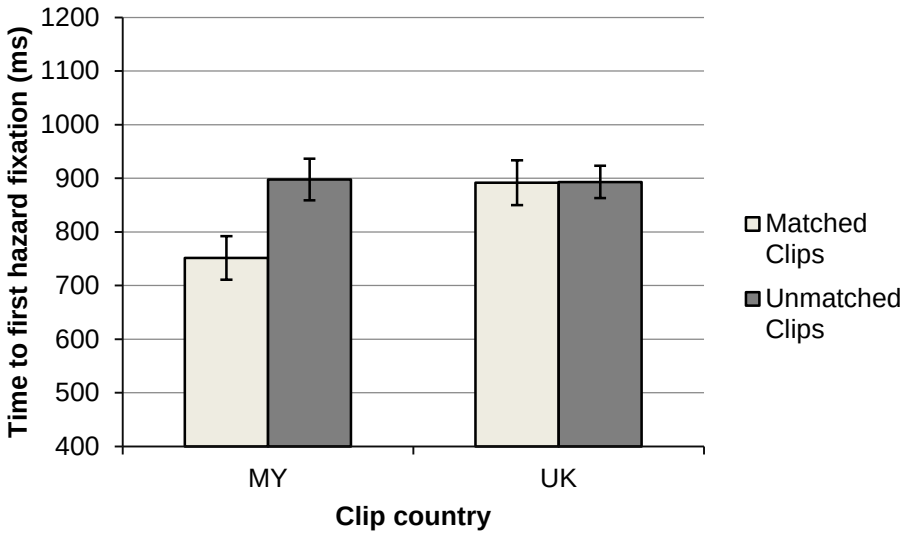

Figure 2-14: Time to first fixate hazards, showing an interaction of matching and clip country. Matched Malaysian hazards were fixated particularly quickly. Error bars represent standard error of the mean.

2.3.2.2 *Mean fixation duration*

Participants' mean fixation duration was calculated for two windows during each clip: first from the start of the clip until hazard onset (*pre-onset*), and secondly from hazard onset to offset, i.e. the hazard window described earlier (*post-onset*), as in Chapman and Underwood (1998). For this particular measure, the ANOVA used in all previous analyses was conducted but with the additional within-groups factor of pre- or post-hazard onset, resulting in a 5-way mixed ANOVA. **Figure 2-15** shows results for all drivers pre-onset and **Figure 2-16** shows results for all drivers post-onset.

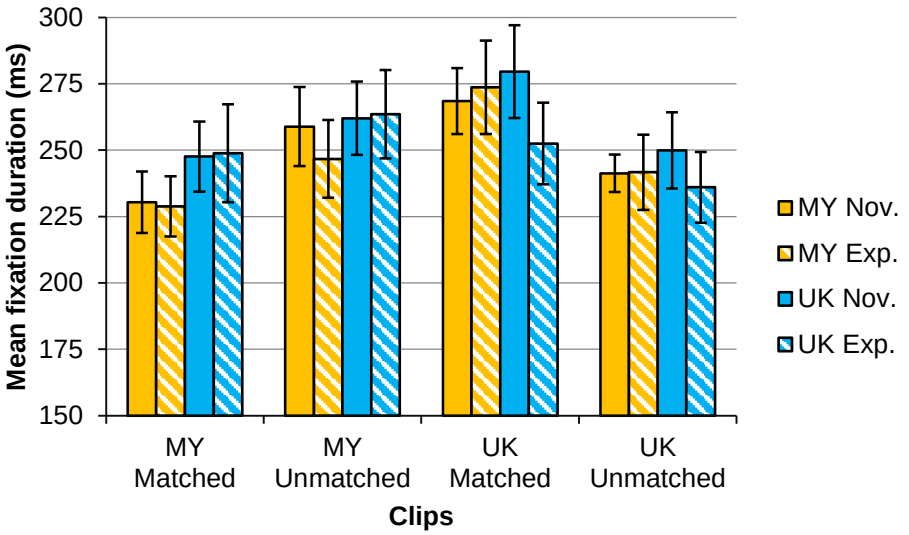

Figure 2-15: Summary data. Mean fixation duration pre-hazard onset. Error bars represent standard error of the mean.

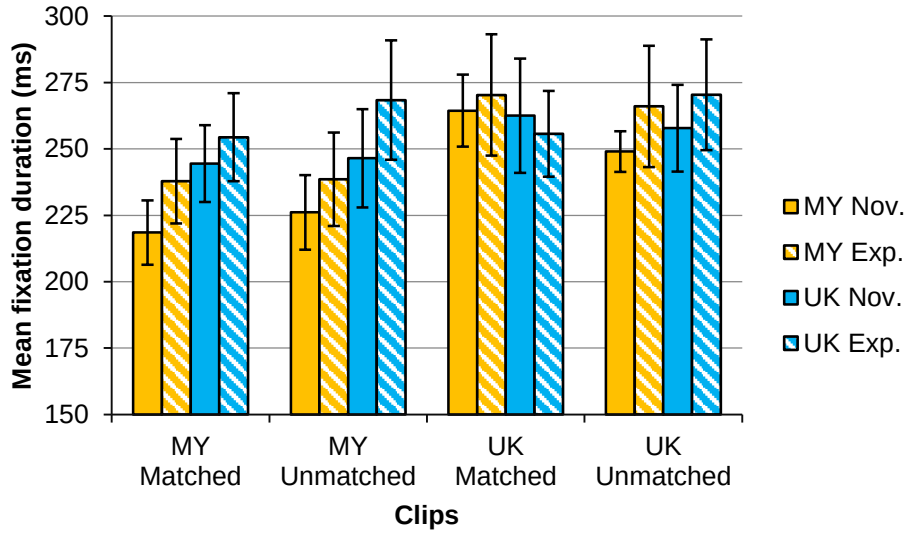

Figure 2-16: Summary data. Mean fixation duration post-hazard onset. Error bars represent standard error of the mean.

A main effect of clip country was found, with shorter mean fixations during Malaysian videos ($F_{1,64} = 8.97$, $p = .004$, $\eta^2_\rho = .123$). Four two-way interactions were found, two shown in **Figure 2-17**. Panel (a) depicts an interaction of clip country and driver origin ($F_{1,64} = 5.65$, $p = .021$, $\eta^2_\rho = .081$), where Malaysian drivers had particularly short fixations in Malaysian clips compared to UK clips ($t(32) = 4.43$, $p <$

.001, d = 0.41). There was also a crossover interaction of experience and pre-/post-hazard onset ($F_{1,64}$ = 4.71, p = .034, η^2_ρ = .069), shown in panel (b), where novice drivers had longer fixations pre-onset, but experienced drivers showed the opposite pattern; however, post hoc tests revealed no significant differences. The remaining two-way interactions, pre-/post-onset and clip country ($F_{1,64}$ = 10.60, p = .002, η^2_ρ = .142) and matching and clip country ($F_{1,64}$ = 24.29, p < .001, η^2_ρ = .275), were subsumed by a three-way interaction of pre-/post-onset, matching and clip country ($F_{1,64}$ = 18.82, p < .001, η^2_ρ = .227), shown in **Figure 2-18**. There were no pre- or post-onset differences in Malaysia matched or UK matched clips, but participants showed shorter fixations post-onset than pre-onset in Malaysia unmatched clips (t(67) = 2.25, p = .028, d = 0.15) and longer fixations post-onset than pre-onset in UK unmatched clips (t(67) = 3.91, p < .001, d = 0.30).

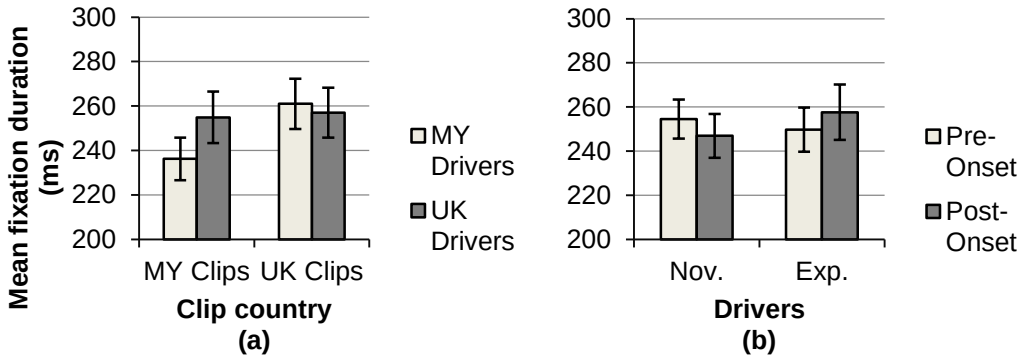

Figure 2-17: Mean fixation duration, showing two 2x2 interactions. Panel (a) displays mean fixation duration across driver origin and clip country; while panel (b) displays mean fixation duration pre-onset and post-onset, across driver experience. Error bars represent standard error of the mean.

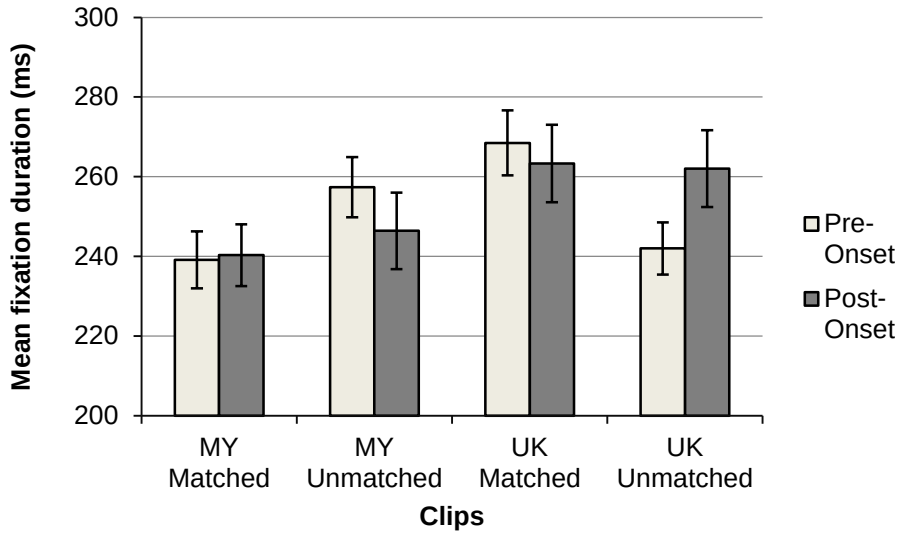

Figure 2-18: Mean fixation duration, showing a 3-way interaction of driver origin, experience level, and matching. Error bars represent standard error of the mean.

2.3.2.3 *Horizontal spread of search*

Analyses for horizontal spread of search were conducted for the period from the start of the clip until hazard onset (the *pre-onset* window used in the above analysis, 2.3.2.2). Data for all drivers is summarized in **Figure 2-19**.

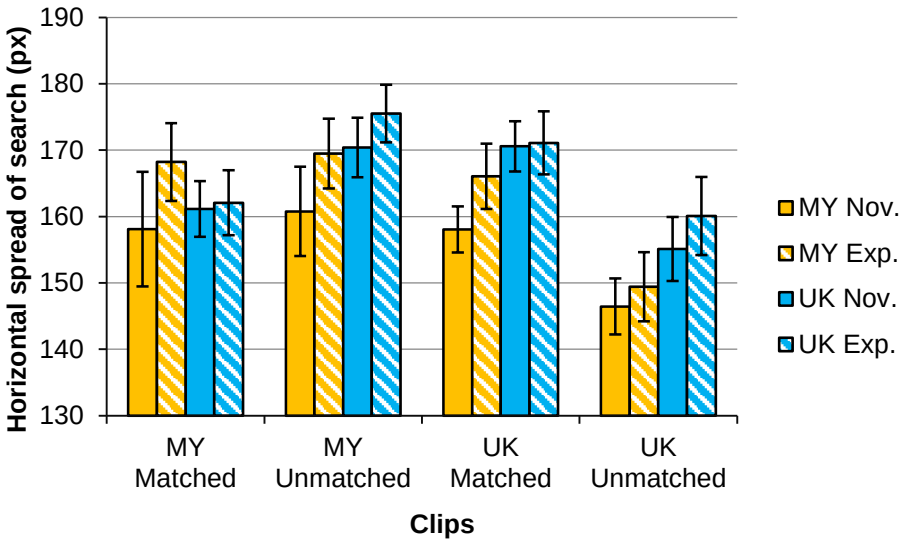

Figure 2-19: Summary data. Horizontal spread of search before hazard onset. Error bars represent standard error of the mean.

There was a main effect of clip country ($F_{1,64} = 8.59$, $p = .005$, $\eta^2_\rho = .118$) and matching ($F_{1,64} = 12.39$, $p = .001$, $\eta^2_\rho = .164$), where there was a wider spread of search in Malaysian and matched clips respectively. There were two two-way interactions; matching and clip country ($F_{1,64} = 111.49$, $p < .001$, $\eta^2_\rho = .635$), and matching and driver origin ($F_{1,64} = 6.68$, $p = .012$, $\eta^2_\rho = .095$). All above effects were subsumed by a three-way interaction of driver origin, matching and clip country ($F_{1,64} = 4.87$, $p = .031$, $\eta^2_\rho = .071$), shown in **Figure 2-20**. Two 2-way ANOVAs (matching and clip country) were conducted for both Malaysian and UK drivers, which were both significant ($F_{1,32} = 34.63$, $p < .001$, $\eta^2_\rho = .520$ and $F_{1,32} = 95.49$, $p < .001$, $\eta^2_\rho = .737$ for Malaysian and UK drivers respectively). Further t-tests revealed that for Malaysian drivers, horizontal search was similar for all clip types except for unmatched UK clips, where spread of search was significantly narrower compared to both unmatched Malaysian clips and matched UK clips ($t(32) = 5.13$, $p < .001$, $d = 0.81$ and $t(32) = 6.77$, $p < .001$, $d = 0.75$ respectively). While UK drivers also had the narrowest horizontal search for unmatched UK clips, unlike Malaysian drivers they

showed a crossover interaction where all possible comparisons were significant. Unmatched clips had narrower horizontal search than matched for clips filmed in the UK ($t(34) = 6.72$, $p < .001$, $d = 0.63$), but the opposite was true for clips filmed in Malaysia ($t(34) = 6.61$, $p < .001$, $d = 0.63$). Similarly, matched Malaysian clips had narrower horizontal search than matched UK clips ($t(34) = 3.81$, $p = .001$, $d = 0.49$), but the opposite was true for unmatched clips, where unmatched UK clips had narrower horizontal search than unmatched Malaysian clips ($t(34) = 5.94$, $p < .001$, $d = 0.74$).

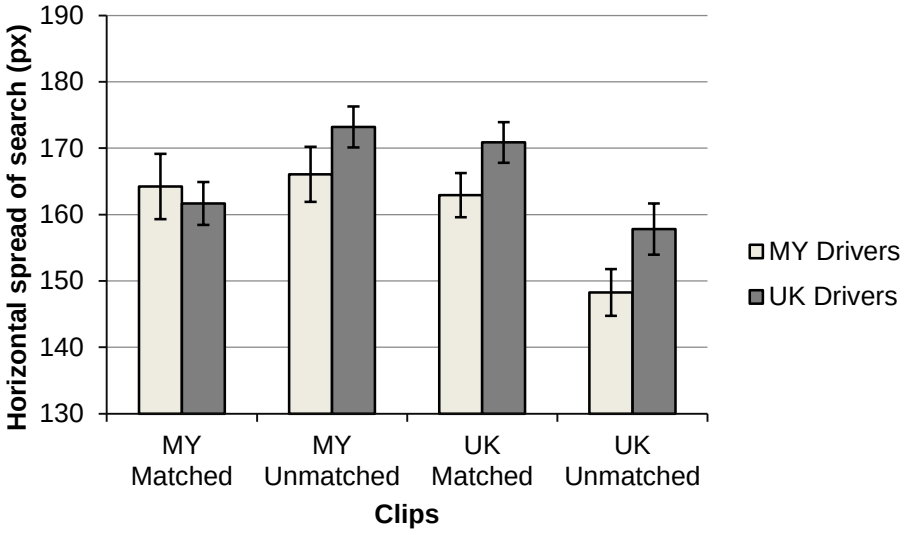

Figure 2-20: Horizontal spread of search, showing a 3-way interaction of driver origin, experience level, and matching. Error bars represent standard error of the mean.

2.4 Discussion

In this discussion, we will focus on how experience and familiarity with both driving environment and hazard types affects drivers' performance. We will also review factors that appear to moderate visual strategies, and the possibility of "look but fail to see" errors (Brown, 2002). Finally, we will discuss how different hazard identification thresholds may affect participants' responses, and the subsequent

implications for hazard perception testing in both developed and developing countries.

2.4.1 Hazard perception skill transferability and the effect of familiarity

In line with Wetton et al. (2010)'s findings, hazard perception skill appears to be highly transferable, as Malaysian and UK drivers showed the same general response pattern for most measures. Differences between drivers were mostly seen when one country's drivers had an exaggerated response to particular clip types without changing the general response pattern; for instance, UK unmatched hazards produced the lowest response rates across all participants, but Malaysian drivers had particularly low response rates on these clips. Performance did vary with all independent variables, however within-groups factors (i.e. clip country and matched/unmatched clips) accounted for more variation than between-groups factors (driver origin and experience), suggesting that while certain nuances of hazard perception are affected by context, drivers view and respond to hazards largely similarly regardless of their home driving environment.

Drivers detected more pre-defined hazards from their own country (although only in the stricter, time-bound analysis) and also hazards that were matched for type. One possible explanation for this is that both environmental and hazard familiarity facilitate drivers' ability to detect hazards in a timely manner. It is also possible that drivers have a lower threshold for identifying hazards when the hazards and/or environment are familiar. However, if this were the case we would also expect to see more extra-hazard responses in home countries. In fact, all participants made more responses, to both pre-defined and extra-hazards, in Malaysian clips, and this tendency was especially pronounced in UK drivers, possibly reflecting a greater general inclination for UK drivers to appraise events as hazardous, and/or the more

hazardous road environment in Malaysia. If, then, we rule out the possibility that drivers have lower thresholds for identifying hazards in a familiar driving environment, we can consider that environmental familiarity facilitates drivers' ability to discriminate between hazards and non-hazards effectively. While Malaysian drivers appear to discriminate equally well in both countries, UK drivers do appear to discriminate more effectively in UK clips. Furthermore, the interaction between driver origin and clip country is only seen in the stricter, time-bound analysis and not the more liberal analysis; given the ability to retrospectively identify hazards with no time pressure, the advantage of environmental familiarity disappears (in fact, it disappears altogether for UK drivers, who detect more pre-defined Malaysian hazards than UK hazards in the more liberal analysis). This suggests that environmental familiarity does indeed prime drivers to react appropriately once a hazard appears, possibly because their mental models are richer in familiar environments; having a greater awareness of the possible hazards and precursors that may occur makes them better equipped to detect early warning signs of dangerous situations (Underwood, Chapman, Bowden, & Crundall, 2002).

It should be noted however that the driver origin / clip country interaction does *not* occur in reaction time analyses, so environmental familiarity does not necessarily mean that drivers react faster, simply that they react within an appropriate timeframe. Perhaps the interaction does not stem from the advantage of a familiar environment, but rather, the disadvantage of an unfamiliar one; it is possible that processing an unfamiliar environment increases cognitive load to the point where drivers may occasionally (but not consistently) fail to discriminate hazards in time.

Hazard familiarity may also affect hazard identification thresholds, although evidence for this is inconclusive. Participants identified fewer unmatched hazards;

this was largely driven by UK unmatched clips, which Malaysian drivers performed particularly poorly on, suggesting that they were less likely to appraise hazards as hazardous when they were unaccustomed to the type of event occurring. Drivers may be more inclined to react to familiar hazards because they are aware of the possible dangers these hazards pose and conversely, less inclined to identify unfamiliar hazards because they lack the necessary experience to appraise them as hazards (Groeger, 2000).

2.4.2 Experience

Contrary to our hypothesis, it appears that driving experience did not have a significant effect on response latency. However, experience did play a role in several interactions, suggesting it may be important cross-culturally but only within specific hazard contexts; for instance, UK experienced drivers appear to be especially sensitive to hazards when both location and hazard type are unfamiliar.

Finding no effect of experience in response rate and reaction time might be explained in Malaysian drivers by the relatively hazardous Malaysian driving environment, compared to the developed countries where previous hazard perception research has been conducted. It is somewhat surprising that there also was no effect of experience among drivers in the UK, where the reaction time paradigm has been found to differentiate experience in the UK and other developed countries (Horswill et al., 2008; Wallis & Horswill, 2007; Wetton et al., 2010). This may be due to an overall lack of experience, as the experienced driver group in this study had an average post-license experience of approximately four years, compared to the above studies where the experienced groups had held their license for over ten years. Indeed, other studies in developed countries have also failed to find significant differences between experienced and novice drivers when the experienced group had

held their license for a relatively short period of time (Chapman & Underwood, 1998; Crundall et al., 2002; Underwood et al., 2013).

However, other studies have also failed to find a difference of experience even with a highly experienced group (Crundall et al., 2003; Sagberg & Bjørnskau, 2006). The lack of differentiation could therefore be for entirely different reasons, such as the nature of the hazards used. For instance, (Sagberg & Bjørnskau, 2006) noted that while their test as a whole did not differentiate experience, certain test items did appear to do so when examined individually. It is also possible that individual differences for identifying hazards may have superseded any difference of experience, particularly among Malaysian drivers; for instance, they appear to have a higher threshold for identifying hazards, and also exhibit greater variance in behavioral metrics compared to UK drivers. This is discussed in further detail in Sections 2.4.4 and 2.4.6.

2.4.3 Visual strategies

Previous research has found that visual strategies vary by both experience and driving environment (Borowsky et al., 2010; Chapman & Underwood, 1998; Crundall, Chapman, France, Underwood, & Phelps, 2005; Crundall & Underwood, 1998; Konstantopoulos, Chapman, & Crundall, 2010; Underwood et al., 2002; Underwood, Chapman, Brocklehurst, Underwood, & Crundall, 2003) and thus we hypothesized differences for both experience level and clip country. However, for the most part, eye tracking metrics did not differ substantially between drivers, and varied mostly between clip country and whether the clips were matched. Again, it is surprising to find no difference in eye movement patterns between novice and experienced drivers, as the studies cited above all found differences in visual strategies. However, as mentioned earlier it is possible that the experienced drivers in

this study had not been driving long enough to have sufficiently developed visual strategies; in all studies cited above the experienced group of drivers had held their license for at least 4.5 years longer than the novice group, often substantially more.

Spread of search along the horizontal meridian was greater in Malaysian clips, possibly reflecting the more cluttered, hazardous road environment in Malaysia; Chapman and Underwood (1998) found similarly increased horizontal search for participants watching videos filmed in urban areas, compared to rural and suburban. Fixations were also shorter in Malaysian videos, again potentially a reflection of the more cluttered environment, similar to Crundall and Underwood (1998)'s findings of decreased fixation durations in demanding roadways. However, this was also driven by Malaysian drivers having particularly short fixations in these clips; their eye movements may have been more efficient in a familiar environment, allowing them to process more visual information in a shorter time (Borowsky, Oron-Gilad, Meir, & Parmet, 2011; Chapman & Underwood, 1998; Crundall et al., 2012; Konstantopoulos et al., 2010). As the UK road environment is less cluttered than Malaysia's and therefore less visually demanding, it is possible that Malaysian drivers adapted their visual strategies to the current environment, similar to the behavior Crundall and Underwood (1998) found in experienced drivers. Interestingly, both novice and experienced UK drivers failed to show this effect and had similar length fixations across both countries' clips.

Participants also had longer fixations upon hazard onset in UK unmatched clips, showing the typical effect of attentional capture (Chapman & Underwood, 1998). Unexpectedly, this was the only clip type where this effect was found; there was no difference in pre- and post-hazard onset in Malaysia and UK matched clips, and in fact the opposite effect was found in Malaysia unmatched clips where fixations

actually became *shorter* upon hazard onset. Furthermore, novice drivers also had shorter fixations upon hazard onset. Both these results are striking as the effect of attentional capture has been consistently found across several studies (Chapman & Underwood, 1998; Chapman et al., 2002; Underwood et al., 2005), with longer fixations during dangerous situations, and novice drivers are particularly susceptible to this effect.

It is possible that fixation length did not increase upon hazard onset in Malaysia because of the nature of the driving environment, which often demands attention be divided among several hazards and possible hazards; hence defining a "danger window" as Chapman and Underwood (1998) did may be redundant, as large parts of the Malaysian videos could be considered as hazardous as Chapman and Underwood's original danger window. The frequency of extra-hazard responses to Malaysian unmatched clips certainly seem to suggest as much. In the UK in contrast, hazards would have been viewed as relatively more hazardous compared to the driving environment, and hence we see a normal effect of attentional capture where fixations increase with hazard onset. However, this does not explain why UK matched videos do not show similar attentional focusing, nor why fixations actually decrease upon hazard onset in the Malaysia matched clips. Further research is required to explore this effect in more depth.

Notably, there were no main effects of driver origin in any of the eye tracking metrics, a sharp contrast from the behavioral results where this effect was found in every metric. This result is interesting in itself, as it suggests that visual strategies are largely moderated by the immediate driving environment, and less so the familiarity of the environment; although both clearly interact. As mentioned, it is possible that the participants in this study had not been driving long enough to have sufficiently

developed visual strategies; if this is the case, there are two possible implications. Firstly, they may not have yet developed strategies specific to their environment (assuming they would at all), and would therefore have more flexible search patterns. Secondly, they might use the same search strategy across all environments (Crundall & Underwood, 1998; Falkmer & Gregersen, 2005), although this seems unlikely as Malaysian drivers did appear to adapt their visual strategies to the UK environment. Regardless, drivers' early visual strategies appear to be similar regardless of where they learned to drive. Furthermore, the lack of driver origin differences implies that Malaysian and UK drivers are using broadly similar visual strategies, and therefore these cannot explain the behavioral differences in Malaysian and UK drivers.

2.4.4 Hazard perception ability and hazard appraisal

Results for response rate and reaction time certainly suggest that UK drivers possess superior hazard perception abilities to Malaysia drivers, as they detected pre-defined hazards in both countries faster and more often. However, other measures such as extra-hazard responses and eye tracking data suggest that different hazard identification thresholds could also be a major contributing factor to performance. This also indicates cross-cultural differences in how dangerous Malaysian and UK drivers perceive hazards to be.

Overall, Malaysian participants appear to have a substantially higher threshold for hazard identification than UK participants: they had a lower response rate to both pre-defined and extra-hazards compared to UK drivers, suggesting that an event (regardless of whether it was actually a hazard) needed to reach a higher level of hazardousness before Malaysian drivers would be willing to identify it as a hazard. However, this also resulted in superior performance in discriminating between highly hazardous and less hazardous events, as Malaysian drivers made fewer extra-hazard

responses for each pre-defined hazard response, compared to UK drivers. These tendencies are likely due to a greater number of hazards in Malaysia and a more hazardous driving environment overall. Wallis and Horswill (2007) also found evidence for a similar response criterion bias affecting hazard perception performance, although they compared novice and experienced drivers rather than drivers of different cultures. If Malaysian drivers do indeed have a higher threshold for hazard identification this could explain at least in part their lower performance in both response rate and reaction times, as hazardous situations would need to progress further before Malaysian drivers would react, if indeed they did at all.

Interestingly, this tendency appears to apply only to hazard identification thresholds, as there were no differences between UK and Malaysian drivers in ratings of hazardousness for hazards they correctly identified. This suggests that although Malaysian drivers may have a decreased tendency to identify an event as a hazard, once it has passed their identification threshold, they view the event similarly to UK drivers.

It is also possible that UK and Malaysian drivers have an altogether different concept of what constitutes a hazard, making the difference one of categorization rather than acceptable thresholds. Indeed, it is debatable as to whether the performance exhibited by UK drivers truly represents superior hazard perception ability, a lower threshold for hazard identification, an altogether different categorization of hazards, or a tendency to over-respond to hazards; clearly, further research is required to separate these effects. However, in practice, a lowered response criterion and tendency to over-respond presumably stem from a greater degree of cautiousness on the road, which results in safer driving overall. The difficulty of whether hazard perception performance reflects actual skill or different

criterion certainly has implications for future tests (see Section 2.4.6); however superior performance likely has much the same effect, regardless of its cause, in actual on-road situations.

2.4.5 "Look but fail to see" in a hazard perception context

An eye tracking analysis also found that when drivers failed to identify a pre-defined hazard, in 96.5% of these cases they still fixated on the hazard during its hazard window at least once: so the vast majority of these non-responses were not due to a failure to visually detect the event, but rather a failure to perceive it as hazardous. We should however distinguish between appraising events as non-hazardous and failing to process them altogether, i.e. a "look but fail to see" error (Brown, 2002). While the latter seems unlikely as the hazards were the most salient events in the clip and participants had only a single task, analysis of fixation length also failed to find the effect of attentional focusing described by Chapman and Underwood (1998) for three of the four clip types, suggesting that drivers may have not recognized the hazardousness of the events in the clips. A further study found that fixating on objects in a driving environment did not necessarily facilitate recall and presumably, processing of them (Underwood, Chapman, Berger, & Crundall, 2003). Evidence therefore supports both possibilities.

2.4.6 Hazard perception: A possible diagnostic tool in Malaysia?

The higher hazard criterion among Malaysian drivers poses a challenge to developing a test that differentiates between experienced and novice drivers, as it is possible that this criterion supersedes any differences of experience that might otherwise be found. Within the present study it is difficult if not impossible to

differentiate actual hazard perception ability from hazard identification thresholds, raising validity issues should the test in its current form be used in Malaysia.

This has wider implications for the export of hazard perception methodologies. If drivers in developing countries with more hazardous road environments appraise hazards similarly to Malaysian drivers (i.e. have higher thresholds for what constitutes a hazard), this tendency will confound hazard perception latencies when applying the reaction time paradigm cross-culturally. It does appear that experience can be differentiated among Malaysian drivers, as experience did play a part in certain measures; however, while the issue of differing hazard thresholds remains, a reaction time paradigm may not be suitable as a diagnostic tool in Malaysia or other countries with similarly hazardous driving environments. Furthermore, while the response bias found in the present study is particularly relevant in developing countries, Wallis & Horswill (2007)'s results indicate that this bias may also be present in developed countries, which carries implications for not only future but also current hazard perception tests.

It is clear that a test is needed where performance cannot be confounded by hazard identification thresholds. Wetton et al. (2010) have suggested breaking down hazard perception into three components: hazard detection, hazard judgment (in terms of interception trajectories) and hazard classification, mostly described as identification thresholds in this chapter (see Section 1.1.4). A number of tests examining only detection and/or judgment have previously been studied and been found to differentiate between novice and experienced drivers in developed countries (see Section 1.1.4 for more detail); for instance, the Deceleration Detection Flicker Test (Crundall, 2009), which has already been applied in Malaysia with mixed results (Lee, Sheppard, & Crundall, 2011), and the "What Happens Next?" test (Jackson et

al., 2009), where participants are asked to predict the hazard before it actually occurs. In the next chapter, we will be focusing on the “What Happens Next?” test.

CHAPTER 3

A PREDICTIVE HAZARD PERCEPTION TASK: THE "WHAT HAPPENS NEXT?" TEST IN MALAYSIA

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

Abstract

Hazard perception (HP) tests are used in several developed countries as part of the driver licensing process, where they are believed to have improved road safety; however, relatively little HP research has been conducted in developing countries, which account for 80% of the world's road fatalities. Previous research suggests that drivers in these countries may be desensitized to hazardous road situations and thus have increased response latencies to hazards, creating validity issues with the typical HP reaction time paradigm. The present study compared Malaysian and UK drivers' HP skills when watching video clips filmed in both countries, using a predictive paradigm where hazard criterion could not affect performance. Clips filmed in the UK successfully differentiated experience in participants from both countries, however there was no such differentiation in the Malaysian set of videos. Malaysian drivers also predicted hazards less accurately overall, indicating that exposure to a greater number of hazards on Malaysian roads did not have a positive effect on

participants' predictive hazard perception skill. Nonetheless the experiential discrimination noted in this predictive paradigm may provide a practical alternative for hazard perception testing in developing countries.

3.1 Introduction

In Chapter 2, we conducted a cross-cultural hazard perception test, using the traditional reaction time paradigm, and found that Malaysian drivers required a higher threshold of danger than UK drivers before they would identify a situation as hazardous. This tendency makes it difficult if not impossible to differentiate actual hazard perception ability from hazard identification thresholds, raising validity issues when using this particular test paradigm. This is certainly not the first time this issue has been raised, and similar concerns have prompted researchers to explore alternative measures unaffected by response criterion. For instance, (Wetton et al., 2010) proposed that in a traditional reaction time task, three distinct judgments affect responses: hazard detection, trajectory prediction, and hazard classification. They subsequently devised a task that required only detection, although they concluded that this particular task might only be valid for older drivers. Similarly, (Crundall, 2009) developed a change detection paradigm that incorporated both detection and trajectory prediction, and found it to differentiate experience among UK drivers. Other tasks have been used with varying degrees of success (Huestegge et al., 2010; Scialfa et al., 2012; Vlakveld, 2014), although all arguably involve a hazard classification judgment to some degree. (Vlakveld, 2014)

The present chapter employs a predictive paradigm that incorporates hazard detection and trajectory prediction: the "What Happens Next?" test, which has been found to differentiate experience among UK and Spanish drivers (Castro et al., 2014;

Jackson et al., 2009). In both studies, drivers watched video clips containing hazards, but the clips were stopped and occluded immediately prior to hazard onset and drivers were asked to predict the events that might have occurred after this point. Participants responded by providing written answers to three questions after every video: (1) What was the source of the hazard? (2) What was the location of the hazard? (3) What happens next? Jackson et al. (2009) found that participants' accuracy in each of these questions decreased significantly, i.e. they most often correctly identified the hazard source, then its location, then what happens next. More importantly, they also found that experienced drivers predicted events more accurately than novices when all clip information was removed from the screen immediately following the occlusion point. Castro et al. (2014) observed similar results, reporting that learner drivers had lower scores than novice and experienced drivers in clips where hazards eventually occurred. They also used videos which did not contain eventual hazards, with a smaller but still significant effect of experience; experienced drivers outscored novices, and additionally, non-recidivist drivers outscored recidivist drivers who had lost all the points on their driving license.

As the present study used the same video clips as Chapter 2, this provides an opportunity to draw comparisons between test paradigms. More importantly, as discussed above, it offers a measure of hazard perception that is unaffected by response criterion. Drivers are not asked to decide whether or not a hazard has occurred; they are merely asked to predict an event. The findings should help establish whether the cross-cultural differences seen in Chapter 2 were entirely the result of a criterion difference, or also reflect differences in hazard perception skill. The present study also employs multiple choice questions unlike the original predictive paradigm, which used a free response format; this serves to establish a

more viable version of the paradigm for large-scale testing, should it again differentiate experience. We hypothesize that experienced drivers will outperform novices on all clips regardless of where they were filmed, however, as in Chapter 2, we also expect this particular advantage to decrease when drivers view clips filmed in their non-home country. As in Chapter 2, the current study also used eye tracking measures for all participants, which allowed us to gauge whether the "What Happens Next?" test necessitates different visual strategies from the reaction time task.

3.2 Methods

3.2.1 Participants

Forty participants were recruited from the UK and 37 from Malaysia, all of whom held full or learner driving licenses from their respective countries and had normal or corrected-to-normal vision. Participants were split into two further sub-groups consisting of novice and experienced drivers, resulting in four groups in total: 19 UK novice drivers (mean age of 22.9 years and licensing time of 8.25 months, except for three learner drivers who had held their permit for an average of 50.7 months), 21 UK experienced drivers (mean age of 23.3 years and licensing time of 54.9 months), 20 Malaysian novice drivers (mean age of 18.0 years and licensing time of 4.5 months) and 17 Malaysian experienced drivers (mean age of 22.5 years and licensing time of 55.8 months). Participants received either monetary compensation or course credit, where the latter was applicable.

3.2.2 Stimuli and apparatus

The original stimuli were the same videos used in Chapter 2, consisting of 20 clips from Malaysia and 20 from the UK, each containing one hazardous event. Examples of these video clips can be seen in **Figure 3-1** and **Figure 3-2**. Each clip

was edited to end immediately prior to hazard onset, while giving enough predictive information for a viewer to deduce or make an intelligent guess as to what would happen next (Jackson et al., 2009). The resulting clips ranged from 2.7 to 43.7 seconds in length. After each clip ended, a black screen was displayed for one second. Four numbered options then appeared on the screen describing four different possible scenarios that could have occurred after the occlusion point, one of which had actually taken place.

The four options for each video were determined via discussion between one Malaysian researcher and one UK researcher, both of whom held a full driving license in their respective countries. Each set of four options was different and unique to the video, and each option represented an event that could have feasibly taken place after the occlusion point. The options were listed in complete sentences, but contained three basic components: the hazard (e.g. "blue car"), its location ("in left lane") and the event that occurred ("pulls into your lane"). In almost all cases the options within one clip differed by at least two of these components.

To ensure that it was not possible to guess the correct scenario from the text alone, eight volunteers, three from the UK and five from Malaysia, were given the 44 sets of scenario options (40 main clips and 4 practice clips) and asked to guess the correct answer without watching the videos. Malaysian volunteers were also asked whether they had any difficulty understanding the scenarios described, to ensure that the options were accessible to non-native English speakers. All volunteers scored at or below chance in this exercise, indicating it was not possible to guess the correct answer without watching the corresponding videos. None of the volunteers in this exercise participated in the later experiment.

Participants viewed the stimuli on a Tobii T60 and T1750 eye tracker under the same conditions as in Chapter 2, Section 2.2.

Figure 3-1: Practice video for the Malaysian block of clips.

Figure 3-2: Practice video for the UK block of clips.

3.2.3 Design

A 2 x 2 x 2 mixed design was used. The between-groups factors were the origin country of the participant (driver origin: Malaysia or UK) and experience level (novice or experienced). The within-groups factor was the country where the clip was filmed (clip country: Malaysia or UK). Matching was not analyzed as a factor in the present or following chapters. Since the large number of Chapter 2 measures allowed a detailed analysis of matched and unmatched clips, in the interest of simplicity matching has been excluded as a factor from all studies onward in order to focus on areas of more applied interest such as driver experience and culture.

As in Chapter 2, the stimuli were separated by country into two blocks of 20 clips, i.e. one Malaysia block and one UK block. Within each block, the order of

clips to be presented was randomized using a Latin square. The order of the blocks was counterbalanced across participants.

Hazard precursors, defined as a foreshadowing element that provides a cue to the hazard (Crundall et al., 2012), and precursor windows were identified for each hazard. A single researcher defined all forty precursors and windows, of which a sub-sample were confirmed by a second researcher with a high degree of agreement. The window began when the precursor to the hazard was first visible and could be considered a reasonable foreshadowing cue (e.g. a vehicle signaling before changing lanes would not have been considered a precursor until the driver began signaling), and ended when the hazard itself was detectable by the viewer, which generally but not always corresponded to the beginning of the hazard window described in Chapter 2 (Section 2.2.4). **Figure 3-3** below shows an example.

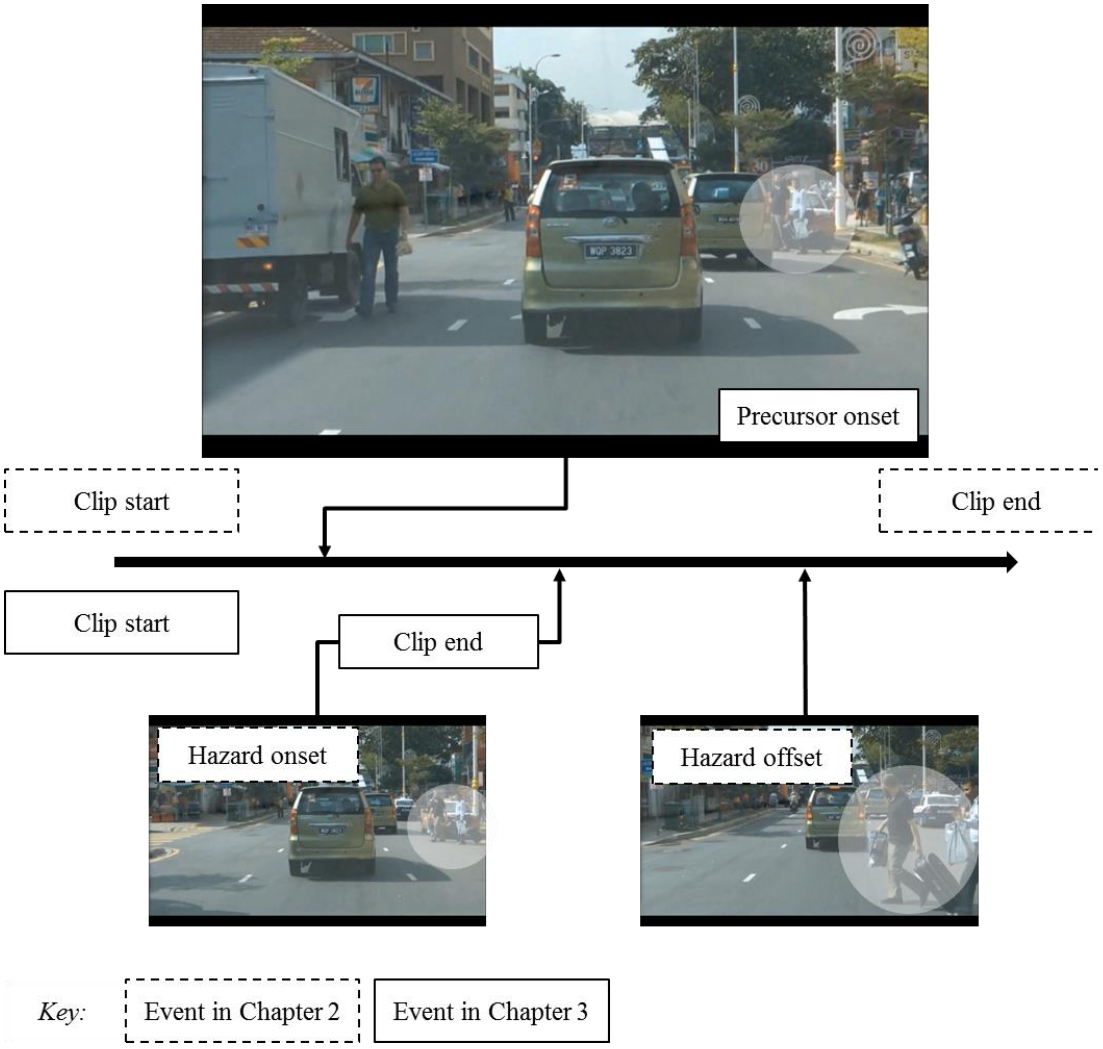

Figure 3-3: Video stills from a sample Malaysian clip illustrating the precursor window. In the example, the precursor window begins as soon as the pedestrians can be seen standing on the side of the road, and ends at the same point the video ends, when the pedestrians begin moving. This is usually the same point as the beginning of the hazard window in Chapter 2.

3.2.4 Procedure

After giving informed consent, participants completed a brief demographic questionnaire and were seated in front of the eye tracker. Participants were informed that each clip contained a driving scenario leading up to a hazardous event, however the clips would end immediately before this event actually occurred and their task was to predict what the event was by selecting the correct scenario out of four possible options. They were informed that in every case, one and only one of the four

scenarios had actually taken place and there was therefore a correct answer for each clip. It was also emphasized that their task was not to choose the event that they felt was the most hazardous, but the one that was most likely to have occurred.

Before starting each block of clips, participants were calibrated on the eye tracker, using a standard 9-point calibration procedure. They then attempted two practice clips, filmed in the same country as the block they were about to view. These practice clips were not used in the subsequent experiment and were the same for all participants. After the practice clips they were able to ask questions or seek clarification. Participants were not given any feedback as to the correct scenarios at any point during the practice clips or main experiment.

A short line of text was displayed for 1 second before each clip, indicating participants' progress through the block. After watching each clip, participants selected the scenario they thought most likely to occur by pressing the corresponding number on a numeric keypad (1, 2, 3, or 4). They were then asked to rate how confident they were in their answer and how hazardous the situation was, on a 6-point scale where a higher rating indicated higher confidence or hazardousness respectively. There was no time limit imposed for participants to answer any of the three questions, and they were able to ask the researcher questions to clarify their understanding of the scenarios. After confirming their third and final answer, the progress text appeared to signal the beginning of the next clip (or end of the block, if appropriate), and the process was repeated until the end of the block. After the first block, participants were given the opportunity to take a brief break, and the process was repeated.

3.3 Results

Accuracy scores were analyzed using a 2 x 2 x 2 mixed ANOVA. The between-groups factors were the origin country of the participant (driver origin:

Malaysia or UK) and experience level (novice or experienced). The within-groups factor was the country where the clip was filmed (clip country: Malaysia or UK). The relationship between licensing time, accuracy scores, answer confidence, perceived hazardousness, and three self-reported measures (driving ability, awareness of other road users, and general driving confidence) was then examined. Chi-square goodness of fit tests were then conducted for individual clips to analyze the plausibility of the incorrect, distractor options in each video. Finally, three eye tracking analyses were conducted: spread of search along the horizontal meridian, mean fixation duration, and time spent fixating on precursors in relation to accuracy.

3.3.1 Behavioral analyses

3.3.1.1 *Accuracy*

Accuracy results are summarized in **Figure 3-4**. 0.5% of participant responses were deemed invalid due to incorrect keypresses and excluded. Main effects were found for all three factors: driver origin, where UK drivers outscored Malaysian drivers ($F_{1,73} = 7.58$, $p = .007$, $\eta^2_\rho = .094$); driver experience, where experienced drivers outscored novices ($F_{1,73} = 4.38$, $p = .040$, $\eta^2_\rho = .057$); and clip country, where participants were more accurate on Malaysian clips ($F_{1,73} = 25.98$, $p < .001$, $\eta^2_\rho = .262$). The latter two effects were subsumed by an interaction of experience and clip country ($F_{1,73} = 7.89$, $p = .006$, $\eta^2_\rho = .098$), driven by novices performing particularly poorly on UK clips. Post hoc analyses revealed that experienced drivers outscored novices only on UK clips ($t(75) = 3.12$, $p = .003$, $d = 0.72$) and there was no difference of experience in Malaysian clips ($t(75) = .365$, $p = .716$, $d = 0.08$). Furthermore, experienced drivers' accuracy was similar on both sets of clips ($t(37) =$

1.57, p = .125, d = 0.29), but novices were significantly better at predicting events in Malaysian clips compared to UK clips ($t(38)$ = 5.40, p < .001, d = 0.87).

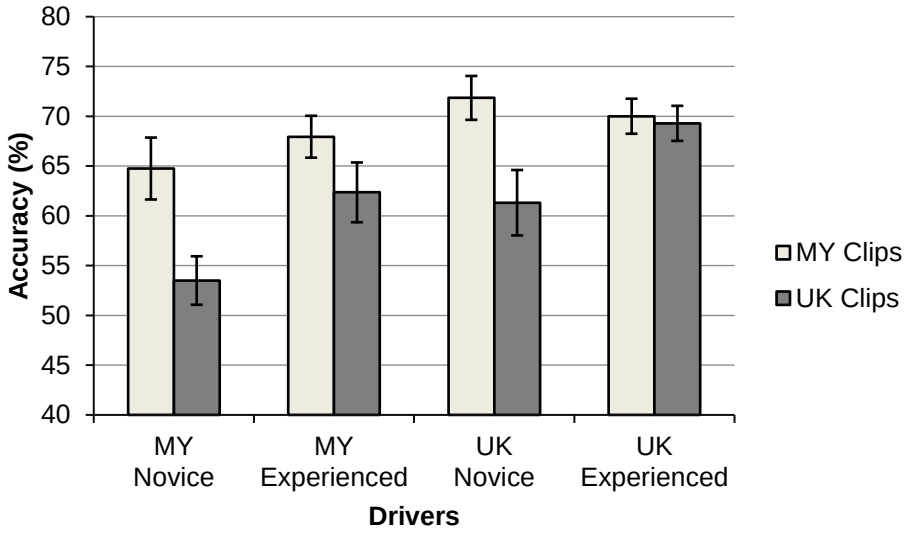

Figure 3-4: "What Happens Next?" scores, based on accurate predictions. Error bars represent standard error of the mean.

3.3.1.2 *Correlational analyses*

Correlations were conducted across all participants to assess whether predictive accuracy was related to a number of factors, including licensing time, two further experimental measures (the hazard and confidence ratings that participants gave for each clip), and three self-reported measures (driving ability, awareness of other road users, and general driving confidence). As there were 21 correlations in total, the false discovery rate (FDR) method was used to adjust the α-level to .009. **Table 3-1** reports the results.

As expected, all three self-rated measures were strongly correlated (all ps < .001 and all rs > .500), suggesting that participants tended to rate themselves similarly on all three measures. However, only driving ability and awareness of other road users marginally correlated with participants' accuracy scores, and this was not significant (r = .270, p = .017 and r = .253, p = .026 respectively). Self-rated driving

confidence correlated with the experimental measure of participants' confidence in their answers (r = .358, p = .001), suggesting that participants appeared to exhibit similar levels of confidence in both their driving ability and answers in the clips.

Similar to (Jackson et al., 2009), participants that rated clips as more hazardous were also more confident in their answers (r = .484, p < .001) . As also observed by (Jackson et al., 2009), there was no relationship between participants' answer accuracy and confidence (r = .024, p = .833).

Finally, licensing time was linked with only driving ability out of the three self-rated measures (r = .300, p = .009), and also marginally correlated with accuracy (r = .220, p = .058), although this was not significant.

Table 3-1: Correlations for all drivers.

	Accuracy score	Hazard rating (clips)	Confidence in answers (clips)	Driving ability	Awareness of others	Confidence in driving
Licensing time	.220	-.179	-.019	.300*	-.080	.118
Accuracy score	-	-.016	.024	.270	.253	.029
Hazard rating (clips)	-	-	.484*	.073	.038	.030
Confidence in answers (clips)	-	-	-	.152	.057	.358*
Driving ability	-	-	-	-	.560*	.720*
Awareness of others	-	-	-	-	-	.504*

*Significant at FDR-corrected α = .009 (two-tailed)

3.3.1.3 *Distractor option plausibility*

To determine whether the distractor options (i.e. the three incorrect options) were equally plausible, goodness of fit tests were performed on individual clips. Only incorrect options chosen by participants were included in this analysis; correct responses were excluded. Ten clips were analyzed with an exact multinominal test due to having particularly low sample sizes (<5 expected responses in each cell), and a chi-square goodness of fit was conducted on the remaining 30. An FDR-corrected α-value of .0224 was used to determine significance. Results are reported in **Table 3-2** and **Table 3-3**.

Out of 40 clips, participants' incorrect answers were not equally distributed in 17 (10 Malaysian clips; 7 UK clips), suggesting that for these clips, one or more of the distractor options was chosen substantially more often compared to the others. To

ascertain any relationship between the distribution of the distractor options and how well a clip differentiated driving experience, effect sizes (given by Cohen's *d*) of the novice/experience difference were calculated for each clip. A Pearson's *r* correlation was then conducted using the chi-square value obtained above and its novice/experience effect size. The correlation was not significant ($r = -.258$, $p = .108$).

Table 3-2: Response distribution for Malaysian clips. χ^2 analysis conducted for only distractor options.

Clip	Correct response	Distractor 1	Distractor 2	Distractor 3	χ^2	*p*
MY-M-01	47	11	0	19	18.20	<.001*
MY-M-02	56	14	3	3	12.10	.002*
MY-M-03	23	42	5	6	50.30	<.001*
MY-M-04	72	1	2	2	-	1
MY-M-05	25	8	20	23	7.41	.025
MY-M-06	46	18	6	7	8.58	.014*
MY-M-07	75	0	0	2	-	.333
MY-M-08	41	8	6	21	11.37	.003*
MY-M-09	66	6	2	2	-	.371
MY-M-10	60	7	9	1	6.12	.0470
MY-U-01	23	30	17	7	14.78	<.001*
MY-U-02	71	2	1	3	-	.877
MY-U-03	54	5	0	18	22.52	<.001*
MY-U-04	74	2	0	1	-	.778
MY-U-05	57	9	10	1	7.30	.026
MY-U-06	48	23	6	0	29.45	<.001*
MY-U-07	72	0	5	0	-	.012*
MY-U-08	65	3	4	4	-	1
MY-U-09	53	3	9	12	5.25	.072
MY-U-10	29	39	5	3	52.26	<.001*

* Significant at FDR-corrected $\alpha = .0224$

Table 3-3: Response distribution for UK clips. χ^2 analysis conducted for only distractor options.

Clip	Correct response	Distractor 1	Distractor 2	Distractor 3	χ^2	*p*
UK-M-01	56	4	5	3	-	.935
UK-M-02	47	1	14	14	11.66	.003*
UK-M-03	52	3	16	5	12.25	.002*
UK-M-04	49	7	9	11	0.89	.641
UK-M-05	65	4	4	4	-	1
UK-M-06	66	7	0	3	-	.022
UK-M-07	49	11	5	11	2.67	.264
UK-M-08	38	18	6	15	6.00	.050
UK-M-09	16	32	16	12	11.20	.004*
UK-M-10	53	6	13	5	4.75	.093
UK-U-01	61	5	10	1	7.63	.022*
UK-U-02	39	11	3	24	17.74	<.001*
UK-U-03	23	19	15	20	0.78	.678
UK-U-04	36	23	16	1	18.95	<.001*
UK-U-05	50	13	3	10	6.08	.048
UK-U-06	62	3	12	0	15.60	<.001*
UK-U-07	37	17	10	13	1.85	.397
UK-U-08	60	10	3	4	5.06	.080
UK-U-09	34	7	15	21	6.88	.032
UK-U-10	47	9	9	11	0.28	.871

* Significant at FDR-corrected $\alpha = .0224$

3.3.2 Eye tracking analyses

Eye movement analyses were conducted using Tobii Studio's Dynamic Area of Interest (AOI) tool. Using the same criterion as in Chapter 2, 23 participants were removed due to missing eye tracking data, leaving 54 participants in total (15 UK novices; 13 Malaysian novices; 18 UK experienced; 8 Malaysian experienced). The accuracy analysis (Section 3.3.1.1) was re-run with these participants to confirm that the results followed the same pattern as previously.

As noted in Sections 3.3.2.2 and 3.3.2.1, spread of search and mean fixation duration were calculated for the entirety of each video, which is roughly analogous to the pre-hazard onset window described in Chapter 2 (2.2.4; 2.3.2.2; 0).

3.3.2.1 *Mean fixation duration*

Participants' mean fixation duration was calculated for the entirety of each video. Results are summarized in **Figure 3-5**. A two-way interaction of clip country and driver experience was found ($F_{1,50} = 4.31$, $p = .043$, $\eta^2_\rho = .079$), which was driven by experienced drivers having particularly short fixation durations while watching Malaysian clips. Post hoc t-tests revealed that among experienced drivers, fixation durations were significantly shorter for Malaysian clips than UK clips ($t(25) = 3.91$, $p = .001$, $d = 0.39$), but no other differences were found.

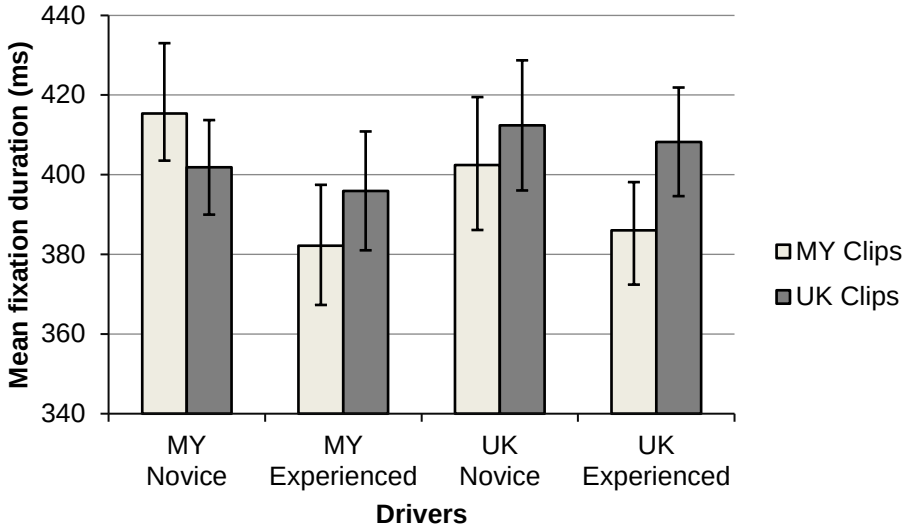

Figure 3-5: Mean fixation duration for the entire clip. Error bars represent standard error of the mean.

3.3.2.2 *Horizontal spread of search*

Spread of search along the horizontal meridian was analyzed for the entirety of each video and is summarized in **Figure 3-6**. There was a main effect of driver origin ($F_{1,50} = 6.72$, $p = .012$, $\eta^2_\rho = .118$), where UK drivers had a wider spread of search than Malaysian drivers. All other main effects and interactions were non-significant.

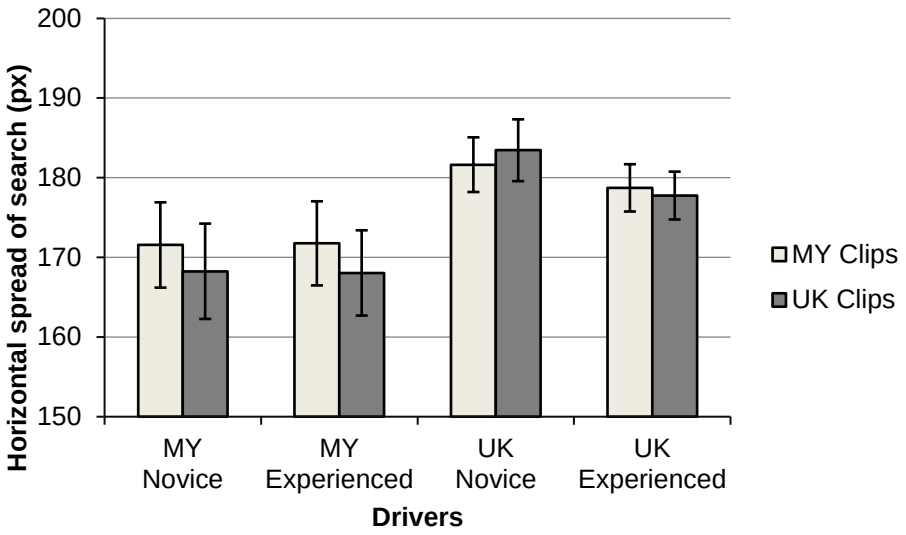

Figure 3-6: Horizontal spread of search for the entire clip. Error bars represent standard error of the mean.

3.3.2.3 *Time spent fixating precursors*

Fifteen videos were excluded from this analysis because less than half of the participants fixated the precursor. Participants' total fixation durations on precursors for the remaining 25 hazards were calculated and z-scored by video. Two fixation duration z-scores were calculated for each participant; for hazards they had correctly predicted and hazards they had incorrectly predicted. Precursors without a fixation were excluded from these scores. A two-tailed paired t-test then compared fixation duration for correct and incorrect videos, and found that participants spent significantly more time fixating precursors for hazards they later predicted correctly (M = .055, SD = .278), compared to hazards they later predicted incorrectly (M = -.161, SD = .581; $t(53) = 2.48$, $p = .016$, $d = 0.473$).

3.4 Discussion

3.4.1 Experience differentiation

Unlike the reaction time paradigm employed in Chapter 2, the "What Happens Next?" test did differentiate between experienced and novice drivers, albeit only for the UK set of clips. Furthermore, while both paradigms utilized the same set of videos, only the predictive task differentiated experience; this is especially notable given that the videos were selected specifically for the reaction time task, raising the possibility that the predictive paradigm may be a more powerful differentiator of experience. Indeed, (Jackson et al., 2009) argue that the "What Happens Next?" test affords a more discerning accuracy measure than response latencies can provide. For instance, a successful response in a reaction time paradigm entails detecting a hazard in an early stage of development, while Jackson et al. (2009)'s paradigm required not only early detection but also future prediction, asking separate questions about a hazard's source ("What is the hazard?"), its location ("Where is the hazard?") and future events relating to it ("What happens next?"). Indeed, Jackson et al. (2009) found that a successful early detection did not necessarily entail an accurate later prediction, as drivers' accuracy dropped significantly with each subsequent question. Furthermore, when using a button press paradigm, one cannot be sure that participants are responding to the same hazard defined by the researchers or a different hazard altogether, although this issue has been compensated for in various ways, such as using a touchscreen or mouse to identify hazard locations, or asking drivers to verbally identify the hazard (Chapter 2; Lim et al., 2013; Wetton et al., 2010, 2011).

3.4.2 Cross-cultural differences

In Chapter 2, we found that both novice and experienced Malaysian drivers were considerably slower to react to hazards than UK drivers, and speculated that this was largely due to a difference in hazard criterion (or *hazard classification*; (Wetton et al., 2010), rather than one in hazard perception ability. In a predictive paradigm such as the "What Happens Next?" test, this possibility has been eliminated, as drivers' accuracy at predicting an event should be unaffected by their opinion of its hazardousness.

However, UK drivers still outperformed Malaysian drivers on the "What Happens Next?" test, suggesting that differences in hazard perception ability may in fact exist between the driver groups, and the previous difference in response latency was likely a combination of actual ability, thresholds for danger, and/or different ideas of what constitutes a hazard. It is also possible that UK drivers' superior performance in both tasks stems from greater participation in hazard perception-type experiments and relevant training, as all UK participants would have practiced for and passed the traditional hazard perception test in order to obtain their license.

Furthermore, while we found a cultural interaction in Chapter 2 where drivers identified more hazards from their home country, no such effect was present in this study, suggesting that a predictive paradigm may be less affected by cultural differences. It also implies that drivers' familiarity with a location has a marginal at best effect on their ability to predict hazards. This lends further evidence to hazard perception skill being highly transferable and relatively unaffected by familiarity with an area, similar to the results of Chapter 2 and (Lim et al., 2013; Wetton et al., 2011). It also has implications for drivers in Malaysia and potentially other developing countries, as the results suggest that a hazardous driving environment may negatively

impact one's hazard perception ability, contributing at least in part to higher accident rates and possibly creating a self-perpetuating cycle.

Given these results and the advantages already discussed in Section 3.4.1, the "What Happens Next" test employed in this chapter is particularly compelling for a Malaysian population. For instance, the absence of the cultural interaction found in Chapter 2 potentially holds implications for drivers obtaining a license outside their home country; they may find a reaction time test disproportionately difficult due to cultural differences, rather than their level of hazard perception skill. More importantly however, the predictive task also circumvents any response bias that may exist in participants, one obvious example being the higher thresholds for hazard identification among Malaysian drivers. As demonstrated in Chapter 2, these biases may confound the reaction time paradigm, as participants' responses may reflect not when they first recognize an event as a potential hazard, but rather, when it has progressed to the point that they are willing to identify it as hazardous. This is particularly relevant in developing countries with higher accident rates, where drivers are more likely to be desensitized to hazards. This finding, combined with the results of the present chapter, suggests that the "What Happens Next?" test may be a practical alternative in countries where desensitization is likely to occur.

3.4.3 Visual strategies

Experienced drivers had particularly short fixation durations while watching Malaysian clips compared to UK clips. This may stem from the more cluttered, hazardous road environment found in Malaysia, similar to (Crundall & Underwood, 1998)'s finding of decreased fixation durations in demanding roadways; experienced drivers may have adapted to the visually demanding environment by increasing their glances (and thereby decreasing their length) around the scene. Spread of search

along the horizontal meridian was also greater among UK drivers; while this may reflect the general behavior of experienced drivers (Chapman & Underwood, 1998), we would also expect to see greater horizontal search in the Malaysian clips, but it appears neither UK nor Malaysian (or indeed, novice or experienced) drivers adapted accordingly. The eye tracking data also revealed that when drivers fixated precursors for a greater period of time, they were more likely to correctly predict the hazard that later occurred. This could indicate that once participants were reasonably certain of a precursor developing into the later hazard, they deployed their attention to focus on the area accordingly; alternately, it could also suggest that the more time participants spent processing the precursor, the more likely they were to identify the later hazard.

We can also compare eye tracking results directly between Chapters 2 and 3, as the two measures employed in this chapter were also analyzed in Chapter 2, across the entire clip and over a similar pre-hazard onset period, making them reasonably comparable, although the analysis windows do differ slightly. Interestingly, visual strategies employed by participants appear to differ between tasks, despite the same videos being used. For instance, Chapter 2's reaction time task appears to be characterized by relatively short fixation durations and narrower horizontal search, implying a strategy of rapid, frequent glances in a relatively curtailed area. In contrast, drivers engaged in the "What Happens Next?" test exhibited lengthier fixations and a greater horizontal search, suggesting longer and wider ranging glances.

Overall however, similar to Chapter 2, we observed few differences of experience in the eye tracking data, a sharp contrast from multiple studies that have reported visual strategies varying with driving experience (Borowsky et al., 2010; Chapman & Underwood, 1998; Crundall et al., 2012; Huestegge et al., 2010;

Underwood, Chapman, Brocklehurst, et al., 2003). As was the case in Chapter 2 (see 2.4.3 for further discussion), it is possible that the experienced drivers in this study had not been driving long enough to have sufficiently developed visual strategies. However, in contrast to Chapter 2, experienced and novice drivers exhibited behavioral but not eye tracking differences, raising the interesting possibility of behavioral differences manifesting faster with experience than the corresponding visual strategies. This is particularly interesting given that previous research has found oculomotor and physiological differences but not behavioral (Chapman & Underwood, 1998; Crundall et al., 2003). While at present it is unclear exactly how quickly visual strategies develop with experience, Chapman et al. (2002) examined participants over the course of their first year of driving and found no differences in their control group's visual search patterns at the end of the year. They also found improvements in another group's scanning behavior after training, although analysis suggested that the scanning strategy exhibited by this group was more conscious than the one automatically adopted by experienced drivers, which might indicate that the training affected specific skills related to scanning behavior more than it did general hazard perception skills.

However, it should be noted that once participants with poor eye tracking data were removed, the remaining sample for the predictive task were biased towards UK experienced drivers and particularly lacking in Malaysian experienced drivers. This was not the case for the eye tracking analyses in Chapter 2, where all four groups remained balanced after the data had been cleaned. Any conclusions or comparisons based on the eye tracking analyses in this chapter should therefore be approached with caution.

3.4.4 Experience differentiation in only UK clips

While the "What Happens Next?" test seems promising thus far, one unresolved issue is that only the UK set of clips were found to differentiate experience, while Malaysian clips did not, and this was the case for drivers from both countries. There are several possible explanations for this difference: a ceiling effect may have occurred in the Malaysian clips, there may an inherent difference between the Malaysian and UK driving clips/environment that makes them more or less suitable for this particular paradigm, or the quality of the distractor options may be superior in the UK set of clips. We will examine these possibilities separately.

First, drivers were considerably more accurate on the Malaysian clips (68.6 %) compared to the UK clips (61.7 %). (Jackson et al., 2009) observed a similar effect when using two different conditions, finding that in the more difficult condition, experienced drivers outperformed novices, but there was no group difference in the easier condition. It is therefore possible that a ceiling effect occurred, and the relative ease of the clips meant that there was no difference between experienced and novice driver performance. However, it should also be noted that 68.6% seems rather low to constitute a ceiling effect, especially given that Jackson et al. (2009) observed scores of 80% under similar conditions with a free response paradigm.

There may also be an inherent difference between the UK and Malaysian clips beyond simple ease of prediction, which makes the Malaysian clips less suited for differentiating experience given a multiple choice predictive paradigm. One obvious difference is that the clips used to reflect the driving environment in Malaysia were generally more hazardous and more visually cluttered (see **Figure 3-1** and **Figure 3-2** for a comparison). It may be the case that the Malaysian driving environment necessitates a more even spread of attention compared to driving in the UK; in other

words, a strategy that retains a high level of awareness of the various developing hazards in the environment – and therefore better readies the driver to deal with any of them – while deploying only limited attention to an immediate hazard, may be more conducive in Malaysia.

Finally, the quality of the distractor options for individual hazards may play a role in experience differentiation; of the seventeen clips where incorrect answers were unevenly distributed, ten were filmed in Malaysia while seven were filmed in the UK. This suggests that when all distractor options appear equally plausible, clips may differentiate experience more successfully. This conclusion is somewhat tentative as experience differentiation for clips (as measured by effect size) did not significantly correlate with their chi-square values, and the difference between ten and seven clips is arguably minor. However, it is notable that the correlation was negative (r = -.258), suggesting that if a relationship does indeed exist, it is likely to be that clips with more plausible distractor options are more effective at differentiating experience.

This possibility is particularly relevant given that this same predictive paradigm has differentiated experience successfully, but using a free response rather than multiple choice format. It may be the case that all else equal, a free response format may be more successful in differentiating experience than a multiple choice one, given the difficulties of devising three alternate, equally plausible responses. From a practical standpoint however, a free response test takes considerable time to score; while (Castro et al., 2014; Jackson et al., 2009) successfully employed the free response format as a diagnostic tool, thus far the "What Happens Next?" test has primarily been used for training (Chapman, Van Loon, Trawley, & Crundall, 2007; Poulsen et al., 2010; Wetton et al., 2013), and a multiple choice format is necessary if the paradigm is to be a viable option for widespread testing.

3.4.5 Further research

The following three chapters (4, 5, 6) detail three "What Happens Next?" projects designed as direct follow-ups to the present experiment, which were all conducted simultaneously. Although none of the following study designs were informed by the results of the others, we have chosen the most logical order in which to describe these studies, and will compare and contrast results of all previous chapters as we progress through this thesis. Chapter 4 directly compares the original free response format used by Jackson et al. (2009) and the multiple choice format used in the present chapter. Chapter 5 explores how important a role drivers' explicit knowledge plays in the "What Happens Next?" task, compared to cognitive processes such as attention and perception. Finally, Chapter 6 examines these cognitive processes in more detail by observing drivers' "What Happens Next?" performance under conditions with limited visual information.

CHAPTER 4

COMPARING "WHAT HAPPENS NEXT?" RESPONSE FORMATS

Abstract

Cross-cultural work on hazard perception has suggested that the standard reaction time test may be less effective in developing countries due to decreased response sensitivity. A predictive paradigm, the "What Happens Next?" test, has been suggested as a possible alternative, however it has usually been researched using a free response format, which has limited practical effectiveness because responses need to be interpreted. The present study compared the free response "What Happens Next?" format to a four-option multiple choice test format, and found that the free response format was a more powerful differentiator of experience. While this is certainly due at least in part to its greater statistical power, it also seems possible that highlighting potential precursors in the multiple choice task benefits novices more than it does experienced drivers.

4.1 Introduction

In Chapter 3, we used a predictive paradigm, the "What Happens Next?" test, and found that only the UK set of video clips differentiated driving experience, while the Malaysian set of clips did not. Thus far, the "What Happens Next?" test has been validated with only clips filmed in Spain and the UK (Chapter 3; (Castro et al., 2014; Jackson et al., 2009; Lim et al., 2014), raising the question of whether, as discussed in

the previous chapter, the Malaysian driving environment in general is less suitable for use with this paradigm compared to more developed countries, given its higher unpredictability, visual clutter, and hazard frequency, particularly abrupt-onset hazards. This would also explain in part the poorer performance of Malaysian drivers on the "What Happens Next?" test, as the unpredictable driving environment and frequent, often simultaneous, hazard occurrences in Malaysia may necessitate a more even spread of attention than driving on UK roads might. Notably, this distinction only applies to the environment and not the drivers themselves, as the "What Happens Next?" test appears to be a valid measure of driving ability among Malaysian and UK drivers alike given the results of Chapter 3.

It is also possible that the test used in Chapter 3 may have been a weaker differentiator of experience in general compared to the Jackson et al. (2009) study due to the format of the task, namely its utilization of multiple choice. Although this still does not explain the UK clips differentiating experience when the Malaysian ones did not, it is possible that the Malaysian clips are simply less effective, which results in no experience differentiation when combined with a less powerful test format. Previous "What Happens Next" studies have used only a free response format (Castro et al., 2014; Jackson et al., 2009), which allows a broader range of scores for each hazard/clip and thus, presumably, a more powerful test for differentiation compared to the binary of multiple choice.

Unfortunately, an open response format is more time-consuming to score compared to a multiple choice test because responses must be interpreted, often by multiple individuals. While interesting for research purposes, this makes the free response format much less practical for mass testing; thus far, most researchers have used the "What Happens Next?" test for training rather than a diagnostic tool

(Chapman et al., 2007; Poulsen et al., 2010; Wetton et al., 2013). However, given that the quality of a clip's distractor options may affect how well it differentiates experience, a free response paradigm may still be useful to inform the creation of the distractor options. It may also reveal some insight into the differences between the two response formats, as the tasks are ostensibly similar but could actually involve different skillsets, given that the multiple choice task effectively highlights potential precursors for participants while the free response task does not. Past research indicates that highlighting potential precursors might benefit novices substantially more, because they have greater difficulty identifying potential risks in the driving environment. For instance, Garay-Vega and Fisher (2005) tracked drivers' eye movements in a simulator and reported that novices fixated both potential precursors and subsequent areas of risk less often than experienced drivers. Furthermore, the precursors were specifically designed to foreshadow and highlight areas of risk (for instance, a pedestrian using a crosswalk, part of which was obscured by a parked truck), but novices often failed to make use of these cues; even after fixating a precursor, they were still less likely to later fixate the relevant risk area compared to experienced drivers. In another simulator study, Crundall et al. (2012) found that learner drivers fixated behavioral prediction (BP) precursors – precursors that were the same stimulus as the hazard – less often than more experienced drivers, and were slower to fixate the eventual hazard after its onset. This suggests that while novices might be peripherally aware of certain road users, they may not think of them as precursors and be unaware of the potential danger they pose.

It is therefore possible that novices find it disproportionately harder to generate potential hazards on their own compared to experienced drivers (free response "What Happens Next?"), but are better at choosing the hazard that

eventually develops out of potential precursors that have already been identified for them (multiple choice "What Happens Next?"). While novices still remain at a disadvantage in this second task, the performance gap is somewhat smaller. This possibility is especially interesting because it might partly explain why the set of Malaysian clips in Chapter 3 failed to differentiate experience (see Chapter 3, 3.4.4): hazard precursors in the Malaysian clips are generally more salient compared to those in the UK, and the average Malaysian clip has more precursors than the average UK clip, which often has only one.

The purpose of the present study is therefore twofold: first, to use participants' unrestricted responses to create the distractor options for a multiple choice format, and second, to directly compare the free response and multiple choice paradigms. We conducted two experiments with different participants, using the same set of videos for both. The first experiment employed the free response procedure similar to (Castro et al., 2014; Jackson et al., 2009), while the second experiment employed the multiple choice format similar to Chapter 3, but used participants' responses from the first experiment wherever possible when creating the multiple choice scenario options.

Because this study was conducted with a new set of Malaysian clips, it had the added potential benefit of validating the "What Happens Next?" test with Malaysian videos, and may help establish whether the videos used in Chapter 3 are simply unsuitable for this particular paradigm, or the problem instead lies in the nature of the driving environment. While we expect experienced drivers to outscore novices in both experiments, we hypothesize greater experience differentiation using the free response paradigm. We also expect higher overall scores in the multiple choice paradigm; as mentioned above, novices should benefit substantially more than

experienced drivers from the switch to multiple choice, although experienced drivers should still outscore them. Finally, while correct responses will almost certainly increase in the multiple choice paradigm, given that the free responses will be used to create the multiple choice distracter options, we anticipate response distribution among the incorrect distractor options being roughly similar between paradigms.

4.2 Experiment 1

4.2.1 Methods

4.2.1.1 *Participants*

Forty participants were recruited from Malaysia, all of whom held full or learner driving licenses obtained in Malaysia, and had normal or corrected-to-normal vision. Participants were split into two groups, consisting of 20 novice drivers (mean age of 18.0 years and mean licensing time of 3.6 months) and 20 experienced drivers (mean age of 27.2 years and licensing time of 95.2 months). Participants received chocolate as compensation.

4.2.1.2 *Stimuli and apparatus*

Twenty clips, ranging 5 to 10 seconds in length, were selected and edited from the Malaysian footage described in Chapter 2 (Section 2.2.2), each containing one hazardous event. As in Chapter 3, each clip was edited to end immediately prior to hazard onset, while giving enough predictive information for a viewer to deduce or make an intelligent guess as to what would happen next (Jackson et al., 2009). After each clip ended, a black screen was displayed for one second. The following three questions were then displayed on the screen: (1) “What was the hazard?”; (2) “Where was the hazard?”; (3) “What happens next?”. The stimuli were played on a 60” projector screen at a resolution of 1368 x 768, using PsychoPy 1.77 (Peirce, 2009).

4.2.1.3 *Design*

A 2 x 3 mixed design was used, where the between-groups factor was the experience level of the participants (novice or experienced), and the within-groups factor was the question answered (hazard, location, prediction). The two practice clips were always the same, while the clips used in the main experiment were played in a different, random order every time.

4.2.1.4 *Procedure*

Participants were tested in groups of up to four at pre-arranged times. After giving informed consent, participants completed a brief demographic questionnaire. They were then given an answer booklet and seated approximately 2m from the screen. As in Chapter 3, participants were informed that they were about to watch a series of video clips filmed from a driver's point of view, each of which contained one hazardous event; however the clips would end immediately before this event actually occurred and their task was to predict what the hazard was. After each clip ended, the image cut to a black screen and participants were asked to provide answers to the questions "What was the hazard?" (*hazard*), "Where was the hazard?" (*location*) and "What happens next?" (*prediction*) in their answer booklet. Participants were given up to one minute to write down their responses to all three questions, and were notified by the experimenter when they had ten seconds remaining.

To ensure participants fully understood the procedure and instructions, all participants first watched two practice clips. After writing down their responses, the experimenter then described the hazard that had actually happened and played the full video clip containing the hazard, allowing participants to compare their answer to the actual event. They did not receive any such feedback for the main experiment.

4.2.2 Results

4.2.2.1 *Experiment 1: Accuracy*

Scoring. The same scoring system employed by Jackson et al. (2009) was used. Two points were awarded for each question that was correctly answered, with a score of 1 for a partially correct answer and 0 for a completely incorrect answer. A maximum score of 2 was therefore possible per question, with a maximum total of 6 points per video and a total hazard score of 120. Two researchers independently marked the score sheets and the means of these scores were used in the analysis. A Pearson's correlation was conducted to check inter-rater reliability and found that $r(800) = .815$, $p < .001$.

Analysis. Three clips were excluded from the original 20 as further scrutiny at a later date suggested that multiple potential events might have occurred after the occlusion point. A 2 (experience level: novice or experienced) x 3 (question type: hazard, location, prediction) mixed ANOVA was conducted on the hazard scores for the remaining 17 clips. This ANOVA was re-run for the original 20 clips to confirm that the results followed the same pattern. **Figure 4-1** shows the mean score for each question type for each group. There was a main effect of experience, where experienced drivers outscored novices ($F_{1, 38} = 10.08$, $p = .003$, $\eta^2_\rho = .210$), and an interaction of question type and experience ($F_{1, 76} = 3.74$, $p = .028$, $\eta^2_\rho = .090$). Two one-way ANOVAs were run on the novice and experienced data respectively, and a significant effect of question type was found for novices ($F_{1, 38} = 4.98$, $p = .012$, $\eta^2_\rho = .208$) but not for experienced drivers ($F_{1, 38} = 1.55$, $p = .226$, $\eta^2_\rho = .075$). Paired t-tests revealed that novices had the highest scores on the hazard question compared to both other questions ($t(19) = 3.33$, $p = .004$, $d = 0.48$ compared to location; $t(19) = 2.25$, $p = .037$, $d = 0.36$ compared to prediction), but there was no difference between

the location and prediction questions ($t(19) = -.501$, $p = .622$, $d = 0.08$). Additionally, experienced drivers outscored novices on every question type ($t(38) = 2.47$, $p = .018$, $d = 0.80$; $t(38) = 3.62$, $p = .001$, $d = 1.17$; $t(38) = 2.89$, $p = .006$, $d = 0.94$ for hazard, location, and prediction respectively).

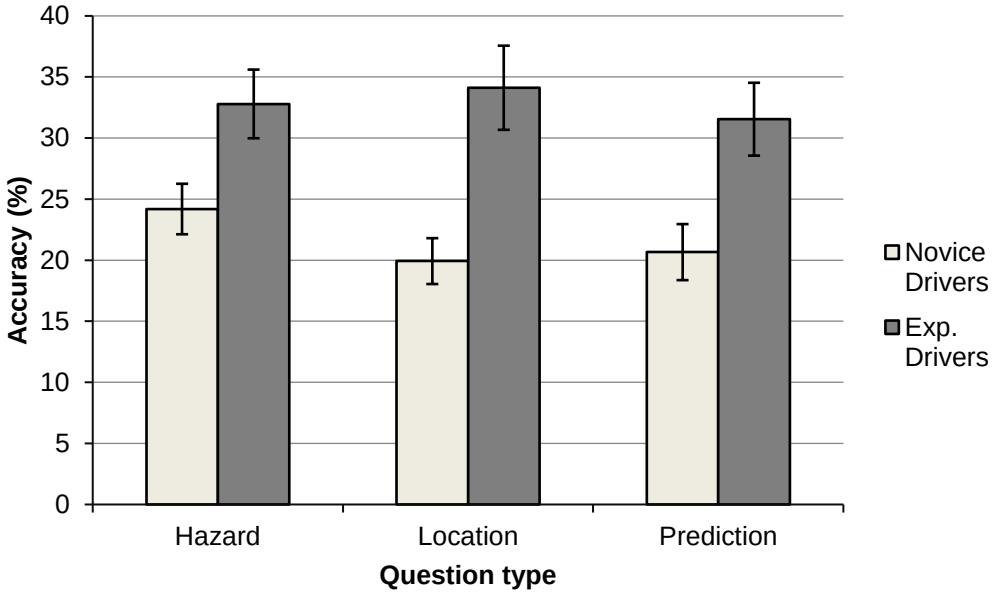

Figure 4-1: Hazard score for the three question types, for both novice and experienced drivers. Error bars represent standard error of the mean.

4.3 Interim discussion

Overall accuracy was extremely low (22.9% for novices and 34.4% for experienced drivers) and somewhat more varied compared to (Jackson et al., 2009), where average scores ranged from 72% to 80%. (Castro et al., 2014) observed scores that were somewhat more comparable but still higher than the present experiment, finding average scores of 34.5% for hazardous situations and 48.6% for quasi-hazardous situations. Notably, both these studies also found main effects of driver experience. This further validates the "What Happens Next?" free response paradigm, as it has now been found to successfully differentiate driving experience with Malaysian, Spanish, and UK drivers in their respective countries.

While the present experiment replicates the overall experience differentiation reported by Jackson et al. (2009), there are several interesting differences otherwise. Lower overall scores are somewhat unsurprising, since Malaysian drivers had lower scores than UK drivers in Chapter 3's "What Happens Next?" experiment, although of course this was using a multiple choice format. However, the novice/experience performance gap in both Chapter 3 and in Jackson et al.'s study was much smaller than the gap between the present experiment and Jackson et al., so this is unlikely to be the whole explanation. The lower scores in the present experiment are most likely due, at least in part, to the videos being more difficult for a predictive task.

The higher video difficulty may also partly explain the two studies' differing pattern of results for question type. Jackson et al. (2009) found a main effect of question type but no interaction, and observed that all drivers' scores significantly dropped with each question, from hazard to location to prediction. This contrasts with the present study, where driving experience interacted with question type: while experienced drivers had similar scores on all three questions, novices found the hazard question easiest but were equally accurate in the location and prediction questions. It seems the videos were challenging to the point that participants had difficulty noticing even the eventual hazard; however, it appears that if they did notice it, it was then relatively easy to then discern its location and what might have happened, although novices still found the latter two tasks relatively difficult compared to experienced drivers.

The different pattern of results for question type may also stem from differences in driver origin, the driving environment in the clips, or some combination thereof. As discussed in previous chapters, the driving environment in Malaysia is more unpredictable and has a higher frequency of hazards compared to the UK,

especially abrupt-onset and simultaneously occurring hazards. This necessitates a more reactive than proactive driving strategy, because drivers often have less forewarning and/or may face several dangerous situations developing simultaneously. It may also mean that drivers in Malaysia spread their attention more evenly, and therefore devote less cognitive resources to any particular situation at any one time compared to drivers in the UK, except for the most salient hazards. This could result in drivers in Malaysia only fully processing the most salient hazards and therefore being able to identify a hazard, its location, and predict what might happen next with similar accuracy. In contrast, drivers in the UK may be able to identify hazards and their location more easily due to less hazards/precursors on the road, but may find it relatively more difficult to predict what might happen next.

It should be noted however that any comparisons between the present findings and Jackson et al.'s are somewhat tentative, due to the contrast in difficulty and cultural settings between the two studies. The responses from Experiment 1 indicate that participants often were unsure as to the hazard, and frequently responded with "no hazard / I did not see anything" or gave a response that was not appropriate for the task; for instance, potentially hazardous events that had occurred earlier in the clip but had clearly passed the point where they might have materialized, or events that were hazardous to another road user but not the camera car. Given the responses to the easier practice videos and the fact participants did respond correctly to a small number of videos, it is unlikely that these responses stem from a failure to understand the task; rather, they are most likely due to participants failing to see the correct hazard and responding with any hazardous event, regardless of how appropriate it may be. While the videos are certainly valid for use in the free response paradigm, since they did differentiate experience, it is somewhat difficult to draw conclusions

about the free response paradigm in Malaysia based on only this set of results, as different sets of videos may produce different patterns of results, at least regarding question type.

Experiment 2 used the same videos as Experiment 1 and a similar procedure, but used the same multiple choice format described in Chapter 3 instead of open responses. Whenever possible, the four options for each video were based on the most frequent responses given in Experiment 1, although due to the low number of usable responses described above, it was usually not possible to create all four options using this method. **Table 4-2** under Results shows a comparison of response formats (described in Section 4.4.2.3) and therefore the number of options carried over from Experiment 1.

4.4 Experiment 2

4.4.1 Methods

4.4.1.1 *Participants*

Thirty-eight participants were recruited from Malaysia, all of whom held full or learner driving licenses obtained in Malaysia, and had normal or corrected-to-normal vision. Participants were split into two groups, consisting of 19 novice drivers (mean age of 18.2 years and licensing time of 8.1 months, except for 5 learner drivers who had held their permit for an average of 4.8 months) and 19 experienced drivers (mean age of 26.7 years and licensing time of 106 months). Participants received chocolate as compensation. None of the participants in this experiment had previously participated in Experiment 1.

4.4.1.2 *Stimuli and apparatus*

The stimuli were the same videos used in Experiment 1, except after each clip ended, instead of the three open-ended questions described in Section 4.2.1.2, four numbered options then appeared on the screen. As in Chapter 3, each option described a different scenario that could have occurred after the occlusion point, one of which had actually taken place. Only the 17 clips that were analyzed in Experiment 1 were used in this experiment.

The four options for each video were determined by participants' responses in Experiment 1, where the four most common and appropriate responses were selected to create the options. If there were less than four common responses for a particular video, two researchers determined the remaining options via discussion. As in Chapter 3, each set of four options was different and unique to the video, and each option represented an event that could have feasibly taken place after the occlusion point. The options were listed in complete sentences, but contained three basic components: the hazard (e.g. "blue car"), its location ("in left lane") and the event that occurred ("pulls into your lane"). In almost all cases the options within one clip differed by at least two of these components.

As in Experiment 1, the stimuli were played on a 60" projector screen at a resolution of 1368 x 768, using PsychoPy 1.80. The order of presentation was randomized every time, as was the order of the four options presented after each clip.

4.4.1.3 *Procedure*

The procedure was the same as for Experiment 1, except for the following details. To account for the multiple choice component, the answer booklet given to participants contained a list of video numbers with the numbers 1, 2, 3, and 4 printed after every video in the list; participants were instructed to pick the option they

thought was correct by circling the corresponding number. Participants were given up to one minute to finalize their decision, although they were informed that if they so wished, they could signal to the experimenter to continue to the next clip instead of waiting the full minute. When there were multiple participants in a session, the experimenter did not proceed to the next clip until all participants had indicated they were satisfied with their answer.

As in Experiment 1, participants also watched two practice videos before beginning the experiment. After choosing their answers, the experimenter informed participants of the correct answer, and played the full video clip containing the hazard. They did not receive any such feedback throughout the main experiment.

4.4.2 Results

4.4.2.1 *Experiment 2: Accuracy*

A two-tailed independent t-test compared accuracy scores for novice (M = 7.63; 44.89%, SD = 1.54; 9.03%) and experienced drivers (M = 8.58; 50.46%, SD = 2.12; 12.45%). The maximum possible score was 17. No significant difference was found ($t(36) = 1.58$, $p = .123$, $d = 0.53$).

4.4.2.2 *Experiment 2: Distractor option plausibility*

To determine whether the distractor options (i.e. the three incorrect options) were equally plausible, goodness of fit tests were performed on individual clips. Only incorrect options chosen by participants were included in this analysis; correct responses were excluded. Five clips were analyzed with an exact multinominal test due to having particularly low sample sizes (<5 expected responses in each cell), and a chi-square goodness of fit was conducted on the remaining 12. An FDR-corrected α-value of .0206 was used to determine significance. Results are reported in **Table 4-1**.

Out of 17 clips, participants' incorrect responses were not equally distributed in 7, suggesting that for these clips, one or more of the distractor options was chosen substantially more often compared to the others. To ascertain any relationship between the distribution of the distractor options and how well a clip differentiated driving experience, t-tests were conducted for each clip comparing novice and experienced drivers' scores. A Pearson's *r* correlation was then conducted using the chi-square value obtained above and the effect size (given by Cohen's *d*) of the t-test. The correlation was not significant ($r = -.101$, $p = .700$).

Table 4-1: Response distribution for individual clips in Experiment 2. χ^2 analysis conducted for only distractor options.

Clip no.	Correct response	Distractor 1	Distractor 2	Distractor 3	χ^2	*p*
01	21	8	6	7	0.29	.867
02	8	29	1	4	41.71	<.001*
03	22	4	7	9	1.90	.387
04	20	11	3	8	4.45	.108
05	14	21	2	5	22.36	<.001*
06	14	7	8	13	2.21	.331
07	6	7	10	19	6.50	.039
08	25	14	3	0	19.18	<.001*
09	27	1	3	11	11.20	.004*
10	39	1	2	0	-	.778
11	21	15	5	1	14.86	.001*
12	7	25	6	4	23.03	<.001*
13	34	1	5	2	-	.296
14	10	19	8	5	10.19	.006*
15	34	4	2	2	-	.744
16	10	17	6	9	6.06	.048
17	29	4	8	1	-	.062

* Significant at FDR-corrected $\alpha = .0206$

4.4.2.3 *Experiment comparison: Response distribution*

To investigate whether response distribution was affected by the change in format from free response to multiple choice, chi-square tests of independence were conducted for each individual clip comparing Experiment 1 and Experiment 2, using

options for each clip and response format as variables. Because not all the responses in Experiment 1 were used as eventual options in Experiment 2, the contingency tables used varied between clips depending on the number of options that had been carried over from Experiment 1, with either a 2 x 2 (two options; 6 videos total), 2 x 3 (three options; 10 videos total), or 2 x 4 (four options; 1 video total) table. An FDR-corrected α-value of .0176 was used to determine significance. Results are reported in **Table 4-2**. Of 17 clips, 6 had a significantly different response distribution between Experiments 1 and 2, suggesting that for these clips, response format considerably affected participants' answers.

Table 4-2: Comparison of response distribution between Experiments 1 and 2.

Clip no.	FR[1] / MC[2]	Option 1[3]	Option 2	Option 3	Option 4	$\chi 2$	*p*
01	FR	9	15	7	-	11.42	.003**
	MC	21	8	1	-		
02	FR	5	23	-	-	0.14	.707
	MC	8	29	-	-		
03	FR	2	2	-	-	2.60	.107
	MC	22	4	-	-		
04	FR	6	27	3	-	17.58	<.001**
	MC	20	8	3	-		
05	FR	2	31	1	-	13.19	.001**
	MC	14	21	5	-		
06	FR	1	1	1	-	0.17	.918
	MC	14	13	8	-		
07*	FR	12	10	9	3	6.40	.094
	MC	6	10	19	7		
08	FR	3	6	-	-	2.85	.091
	MC	25	14	-	-		
09	FR	2	31	-	-	30.88	<.001**
	MC	27	11	-	-		
10	FR	34	1	1	-	1.36	.507
	MC	39	2	0	-		
11	FR	20	14	-	-	0	.967
	MC	21	15	-	-		
12*	FR	-	0	22	9	8.79	.012**
	MC	-	7	25	4		
13	FR	36	2	1	-	0.67	.715
	MC	34	1	2	-		
14	FR	4	7	3	-	0.33	.846
	MC	10	19	5	-		
15	FR	23	7	-	-	4.39	.036
	MC	34	2	-	-		
16	FR	5	7	-	-	1.20	.274
	MC	10	6	-	-		
17	FR	9	9	2	-	9.81	.007**
	MC	29	4	1	-		

[1] Free response; [2] Multiple Choice; [3] Correct answer unless otherwise noted (see below)

* None of the responses given in Experiment 1 for this clip were correct

** Significant at FDR-corrected $\alpha = 0.0176$

4.4.2.4 *Experiment comparison: Experience differentiation*

To measure how well each video differentiated experience in each response format, the effect size of the novice/experienced score difference (measured by Cohen's *d*) was calculated for each video in both Experiment 1 and Experiment 2. A Pearson's *r* correlation was then conducted comparing videos in Experiment 1 and Experiment 2, using each video's effect size The correlation was not significant ($r = -.176$, $p = .498$). All effect sizes are listed in **Table 4-3**.

Table 4-3: Effect size of the novice/experience difference for individual clips, in both free response (Experiment 1) and multiple choice (Experiment 2) formats.

Clip no.	Experiment 1 Free response			Experiment 2 Multiple choice		
	% correct, novices	% correct, exp. drivers	*d*	% correct, novices	% correct, exp. drivers	*d*
01	22.9	39.6	0.46	42.1	52.6	0.21
02	8.3	7.9	-0.02	21.1	15.8	-0.13
03	5.0	12.5	0.43	47.4	57.9	0.21
04	11.7	25.8	0.46	36.8	57.9	0.42
05	3.3	15.0	0.49	26.3	42.1	0.33
06	14.6	22.9	0.44	42.1	26.3	-0.33
07	14.6	32.5	0.62	15.8	10.5	-0.15
08	47.1	45.0	-0.07	42.1	73.7	0.66
09	10.4	21.3	0.38	57.9	68.4	0.21
10	55.8	70.8	0.48	89.5	94.7	0.19
11	15.4	36.3	0.68	42.1	52.6	0.21
12	4.6	12.5	0.62	15.8	21.1	0.13
13	82.5	91.7	0.32	73.7	84.2	0.25
14	2.1	17.1	0.60	36.8	10.5	-0.63
15	30.8	65.8	0.88	84.2	89.5	0.15
16	11.3	20.0	0.28	21.1	26.3	0.12
17	26.7	21.3	-0.15	68.4	73.7	0.11

4.5 Discussion

It is difficult to draw any definite comparisons between Experiment 2 and Chapter 3, given the apparent difference in task difficulty. There was again no correlation between how well a clip differentiated experience and how evenly its

distractor responses were spread, although this is an uncertain conclusion because the novice/experienced performance gap in the present study did not reach significance, whereas in Chapter 3 the UK set of videos differentiated experience and there was still a main effect of experience overall.

As expected, overall scores in Experiment 2 were higher than Experiment 1, and novices benefitted substantially more than experienced drivers by switching from free response to multiple choice formats, as the performance gap between the two groups decreased considerably between experiments; in fact, although experienced drivers still outscored novices in Experiment 2, this did not reach significance. As discussed earlier, this is unsurprising given the greater range of scores and therefore power of the free response format over multiple choice; however, it is disappointing that the novice/experienced performance gap decreased to the point where the multiple choice clips failed to differentiate experience at all. It seems likely that novice drivers found it extremely difficult to detect and predict potential hazards at all in Experiment 1, while experienced drivers also found this task difficult but significantly less so. However, having potential hazards highlighted in Experiment 2 reduced most of the advantage experienced drivers may have had, to the point where their scores were only numerically and not statistically higher than novices'. This likely contributes to the higher scores in Experiment 2; while all participants may have found it difficult to discern potential hazards on their own in Experiment 1, once these potential threats were highlighted in Experiment 2, they were better at deciding which of these eventually developed into hazards.

Interestingly, the switch in format appears to affect clips inconsistently with regards to experience differentiation, as there was no correlation between the relative ability of clips to differentiate experience in both experiments. In other words, there

was no relationship between the clips that best differentiated experience in Experiment 1 and the clips that best differentiated experience in Experiment 2. Although this conclusion is moderated by the difficulty of the clips and Experiment 2's lack of experience differentiation, this suggests that while ostensibly similar, the "What Happens Next?" paradigm may index slightly different skills depending on the response format used. This is somewhat supported by the response distribution analysis, as the change in response format appears to affect response distribution for at least some clips.

While the results of Experiment 1 are encouraging, the lack of experience differentiation in Experiment 2 again raises the question of whether the driving environment in Malaysia lends itself to effective "What Happens Next?" multiple choice clips, especially given that these clips better differentiated experience by simply changing the response format. Again, however, it is difficult to draw any definite conclusions given the higher difficulty of the clips compared to Chapter 3 and Jackson et al.'s study; it is certainly possible that yet another set of Malaysian clips may differentiate driving experience more successfully. It is also possible that the set of Malaysian clips used in Chapter 3 may yet differentiate experience, given the manipulations described in Chapter 6.

CHAPTER 5

ARE VIDEOS NECESSARY IN A HAZARD PERCEPTION TEST?

Abstract

The hazard perception skill of a driver refers to the ability to identify potentially dangerous situations on the road. Drivers with greater levels of experience tend to outperform newer drivers in tests of hazard perception skill, and past research has linked hazard perception test performance to on-road crash rates. Thus far, almost all hazard perception tests use visuals of on-road scenarios, usually either actual footage recorded on the road or computer-generated imagery (CGI), which can be both time-consuming and expensive to generate. The current study explored to what extent videos are necessary in a multiple choice predictive hazard perception task, the "What Happens Next?" test, by asking drivers to predict events on the road both with and without viewing the associated video scenarios. Both novice and experienced drivers were unable to predict the later video events without having watched the videos first, suggesting that knowledge of common road hazards is unlikely to affect performance on the predictive task, and that videos are necessary for this version of the hazard perception test.

5.1 Introduction

In Chapter 2, we found no experience differentiation in a hazard perception reaction time task using both Malaysian and UK videos. In Chapter 3, we found

experience differentiation in the "What Happens Next?" test, where participants predicted what event was most likely to happen next. In all versions of this test participants have either generated their own responses (Chapter 4; (Castro et al., 2014; Jackson et al., 2009) or selected one event out of four possible options (Chapter 3;(Lim et al., 2014). This task is relatively unusual among hazard perception-type tests because having to generate or choose future events potentially allows participants to draw on their prior knowledge of road events more heavily than other tasks (e.g. (Huestegge et al., 2010; Scialfa et al., 2012; Wetton et al., 2010).

However, most of these tasks involve rapid hazard identification, which arguably requires less knowledge of prior road events than the "What Happens Next?" test, which asks participants to predict possible road events. Interestingly, while explicit knowledge may play a role in both variations of the "What Happens Next?" test, the different response formats may emphasize it in slightly different ways (see Chapter 4, Section 4.1). For instance, the free response task requires participants to generate their own ideas of future hazards, where a lack of knowledge of common hazards and situations in which they might occur might hinder their ability to think of possibilities; if participants have rarely encountered a certain type of hazard before, it will be fairly difficult to predict it. The multiple choice component on the other hand involves an element of intelligent guesswork, where prior knowledge might play an even greater role; since participants lose nothing from guessing an answer they are not sure of, a driver could eliminate certain options with knowledge of road events that are more or less likely to occur, narrowing their options more effectively.

If experienced drivers' superior performance in the "What Happens Next?" test relies on prior knowledge of likely road events to a greater extent than other hazard perception-type tasks, this may partly explain the lack of group differences

found in Chapter 3's eye tracking data. Because experienced drivers outscored novices in Chapter 3, one might expect that the experienced group would also exhibit more efficient visual search patterns, especially given that previous research has found clear group differences (Borowsky et al., 2010; Chapman & Underwood, 1998; Chapman et al., 2002; Crundall et al., 2003, 2002). However, eye tracking data in Chapter 3 showed no differences in novice and experienced drivers' visual strategies when performing the task. While this might be caused by a relatively small gap in experience between driver groups, or possibly behavioral differences manifesting before visual strategies develop (see Sections 2.4.3, 3.4.3), Huestegge et al. (2010) and Underwood et al. (2002) found group differences with similarly low licensing time differences, although these were still higher than in Chapter 3. Since the primary task is to predict the correct event out of four, experienced drivers may have more accurate knowledge of events that are more likely to happen on the road and use this to their advantage over novices. Therefore, it's conceivable that at least in part, the experience effect found in Chapter 3 may be driven by different prior knowledge of events that are likely to happen on the roads (regardless of circumstances), as opposed to what has been perceived.

If experienced drivers do outscore novices on the "What Happens Next?" test due to greater knowledge of likely road events, even in part, this could have implications for hazard perception testing. Thus far, almost all hazard perception tests present drivers with videos or stills of road scenarios, either recorded on the road or computer-generated (see Section 1.1.4 for examples). However, both these methods can be time-consuming and/or expensive; for instance, Wetton et al. (2011) reported that after recording 190 hours of road footage, 182 scenes were selected as suitable. We would expect similar results in countries with comparable accident rates, although

it should be noted that Wetton et al. were developing a test to be used in Queensland driver licensing and therefore clip selection would have been particularly stringent. In countries with higher accident rates such as Malaysia, hazards certainly occur more frequently; however, this often presents a different set of problems, such as other potential hazards arising in close temporal proximity to pre-defined hazards for testing (see Chapter 2, Sections 2.3.1.2 and 2.4.1 for examples).

The format of the "What Happens Next?" test allows an opportunity to explore an alternative measure of hazard perception; namely, a hazard perception test that is questionnaire-only. Because drivers are asked to choose one event out of four, doing this without watching the associated video requires participants to rely solely on their prior knowledge of likely road events. If this prior knowledge plays a role on its own, we would expect experienced drivers to outscore novices, although possibly by a smaller margin than in other video and/or still-based hazard perception tests. While this is certainly a consideration, if a test that discriminates experienced and novice drivers could indeed be developed without the use of videos and/or stills, this would significantly reduce preparation time and resources.

The current study therefore presented participants with two different tasks. In Part I, the first task, participants read four brief descriptions of hazardous events that could potentially happen on the road; in other words, the scenarios from the videos used in Chapter 3, but without the corresponding videos. They were asked to pick the event out of the four they felt most likely to happen. In Part II, the same participants completed the multiple choice "What Happens Next?" task with the same videos that were used in Chapter 3 and a similar procedure.

As in Chapters 2 and 3, the present study has a cross-cultural aspect. Drivers from both Malaysia and the UK completed the study, which allowed us to investigate

whether Malaysian and UK drivers differ significantly in their idea of what events are more or less likely to happen on the road in both countries. This is particularly relevant in the Malaysian clips, where the nature and type of hazards vary more than they would in the UK.

If explicit knowledge plays a comparable role to perceptual skill in the "What Happens Next?" test, experienced drivers should outscore novices in both Part I and Part II. We would tentatively expect drivers to have higher scores in Part I for their home country, as some Chapter 2 results indicate that drivers' mental models are richer in familiar environments (Section 2.4.1); similarly, we would expect to find response pattern differences between Malaysian and UK drivers in Part I. In Part II, we expect to find similar results to the "What Happens Next?" test run in Chapter 3; in other words, experienced drivers should outscore novices, but only in the UK set of clips, and UK drivers should outscore Malaysian drivers.

5.2 Methods

5.2.1 Participants

Thirty-seven participants were recruited from the UK and 36 from Malaysia, all of whom held full or learner driving licenses from their respective countries, and had normal or corrected-to-normal vision. Participants were split into two further sub-groups consisting of novice and experienced drivers, resulting in four groups in total: 19 UK novice drivers (mean age of 21.2 years and licensing time of 6.8 months, except for 11 learner drivers who had held their permits for an average of 3.3 months), 18 UK experienced drivers (mean age of 24.8 years and licensing time of 86.1 months), 18 Malaysian novice drivers (mean age of 18.2 years and licensing time of 5.1 months, except for 2 learner drivers who had held their permits for an average

of 3.5 months) and 18 Malaysian experienced drivers (mean age of 25.4 years and licensing time of 88.4 months). Participants received either monetary compensation or course credit, where the latter was applicable. None of the participants in this study had previously participated in either experiment described in Chapters 2 and 3.

5.2.2 Stimuli and apparatus

Part I of the experiment employed only the text options describing four potential hazardous events for all videos used in Chapter 3. Part II employed both videos and their corresponding text options, i.e. the same procedure used in Chapter 3, although without the hazard and confidence ratings. The videos were displayed at a resolution of 1280 x 720. Both parts of the experiment were written in Matlab using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997), and were presented using Matlab R2013a on a 13.3" laptop with the screen set to maximum brightness.

5.2.3 Design

A 2 x 2 x 2 x 2 mixed design was used. The between-groups factors were the origin country of the participant (Malaysia or UK), and experience level (novice or experienced). The within-groups factors were the country where the scenario was based (Malaysia or UK), and whether the driving scenarios consisted of only text (Part I), or both text and video (Part II).

For both Parts I and II, stimuli were separated by country into 2 blocks of 20 trials, i.e. one Malaysia block and one UK block. Except for the practice trials, the order of presentation was randomized within each block, so participants read the Part I scenarios in a different order from which they viewed the corresponding Part II videos. Within each set of four options, the order of presentation was also

randomized. The order of the blocks was counterbalanced across participants, although this was consistent across both parts of the experiment; i.e. if a participant completed the Malaysian block first in Part I, they also completed the Malaysian block first in Part II, and vice versa.

5.2.4 Procedure

After giving informed consent, participants completed a brief demographic questionnaire and were then seated in front of the laptop. Instructions for Part I of the experiment were given first and without disclosing details of Part II. Participants were informed that they would be reading descriptions of possible hazardous events that could happen on the road from a driver's point of view, and they would read about these events in sets of four. They were instructed to select the event of these four that they thought most likely to happen by pressing the corresponding number on the laptop keyboard (1, 2, 3, or 4). As in Chapter 3, it was emphasized that their task was not to choose the event they felt was the most hazardous, but the one they felt was most likely to happen. Participants were also informed that the block of events they were viewing were set in either Malaysia or the UK, and that the road traffic flow in both countries was left-hand, in which traffic keeps to the left side of the road. Before beginning each block, participants were informed whether they would be reading about Malaysian or UK events, and then attempted two practice trials set in the same country as the block they were about to complete. After these trials they were able to ask questions or seek clarification.

After completing Part I, the experimenter explained to participants that the events they had just read about were related to specific traffic scenarios that had actually occurred, and in Part II they would watch videos of these scenarios. As in Chapter 3, participants were informed that each clip contained a driving scenario

leading up to a hazardous event, however the clips would end immediately before this event actually occurred and their task was to predict what the event was by selecting the correct scenario out of four possible options; additionally, these four options were ones they had already read about in Part I of the experiment. They were informed that in every case, one and only one of the four scenarios had actually taken place and there was therefore a correct answer for each clip. As in Part I, at the beginning of each block participants attempted two practice clips set in the same country as the upcoming block, which corresponded to the practice text-only trials they had completed earlier in Part I. Participants were not given any feedback as to the correct scenarios at any point during the practice clips or main experiment.

In both parts a short line of text was displayed before each trial, indicating participants' progress through the block. There was no time limit imposed for participants to respond, and they were able to ask the researcher questions to clarify their understanding of the scenarios.

5.3 Results

5.3.1 Part I performance against chance

One-sample t-tests compared how accurately the four participant sub-groups (novice Malaysian drivers, experienced Malaysian drivers, novice UK drivers, and experienced UK drivers) guessed the later, correct video answers against a chance level of 1 out of 4, or 25%. Malaysian and UK videos were analyzed separately, resulting in eight t-tests in total. An FDR-corrected α-value of .00625 was used to determine significance. Generally, drivers were not able to guess the correct video scenarios at better than chance levels; the only exception was UK novices picking UK events. Results are reported in **Table 5-1**.

Table 5-1: Participants' accuracy in Part I for guessing the later, correct video answers. Mean scores for all participant sub-groups and comparisons to chance performance.

Participant group		Clip country	Mean % correct guesses	*p*
Novice	**MY**	MY	25.8	.743
		UK	26.4	.688
	UK	MY	27.9	.069
		UK	32.6	.005*
Experienced	**MY**	MY	23.8	.470
		UK	25.5	.825
	UK	MY	23.4	.475
		UK	29.0	.052

* Significant at FDR-corrected $\alpha = .00625$

5.3.2 Consistency with video answers

Participants' responses in both Parts I and II were compared to the correct video answers (i.e. accuracy, for Part II) and analyzed using a 2 x 2 x 2 x 2 mixed ANOVA. Between-groups factors were the origin country of the participant (driver origin: Malaysia or UK), and experience level (novice or experienced). Within-groups factors were the country where the clip or text was based (country setting: Malaysia or UK), and whether the scenario consisted of only text (Part I), or both text and video (Part II). **Figure 5-1** summarizes Part I data, and **Figure 5-2** summarizes Part II data.

There was a main effect of driver origin, where UK drivers outscored Malaysian drivers ($F_{1,69} = 18.29$, $p < .001$, $\eta^2_\rho = .210$). There was also a main effect of scenario, where drivers had considerably higher scores in Part II ($F_{1,69} = 562.53$, $p < .001$, $\eta^2_\rho = .891$). Finally, there were two significant two-way interactions and a

third trending, all involving the scenario factor. The first interaction was of scenario and driver experience ($F_{1,69} = 9.29$, $p = .003$, $\eta^2_\rho = .119$), where in Part I, novices and experienced drivers had similar scores (although novices were numerically higher; $t(59.04) = 1.74$, $p = .087$, $d = 0.45$), but in Part II experienced drivers had higher scores ($t(71) = 2.48$, $p = .016$, $d = 0.59$). There was also an interaction of scenario and country setting ($F_{1,69} = 8.12$, $p = .006$, $\eta^2_\rho = .105$), where in Part I, participants' scores for UK scenarios trended higher ($t(72) = 1.95$, $p = .055$, $d = 0.31$), but the opposite was true in Part II as participants performed better on Malaysian clips than UK ($t(72) = 2.52$, $p = .014$, $d = 0.31$). Finally, an interaction of scenario and driver origin trended towards significance ($F_{1,69} = 3.70$, $p = .058$, $\eta^2_\rho = .051$); while UK drivers outperformed Malaysian drivers in both Part I and Part II, this difference was significant only in Part II ($t(71) = 1.78$, $p = .080$, $d = 0.42$ for Part I; $t(59.60) = 3.48$, $p = .001$, $d = 0.90$ for Part II).

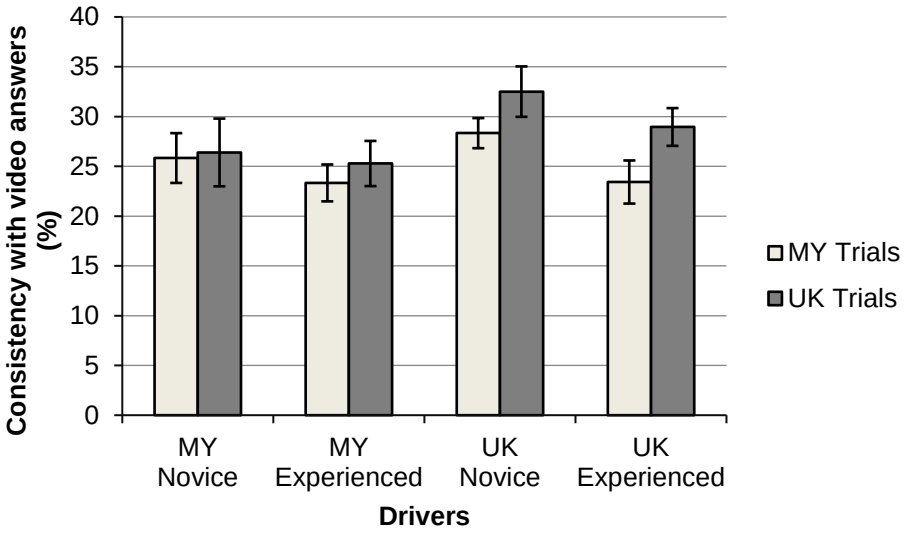

Figure 5-1: Consistency with video answers for all drivers in Part I (text options only). Error bars represent standard error of the mean.

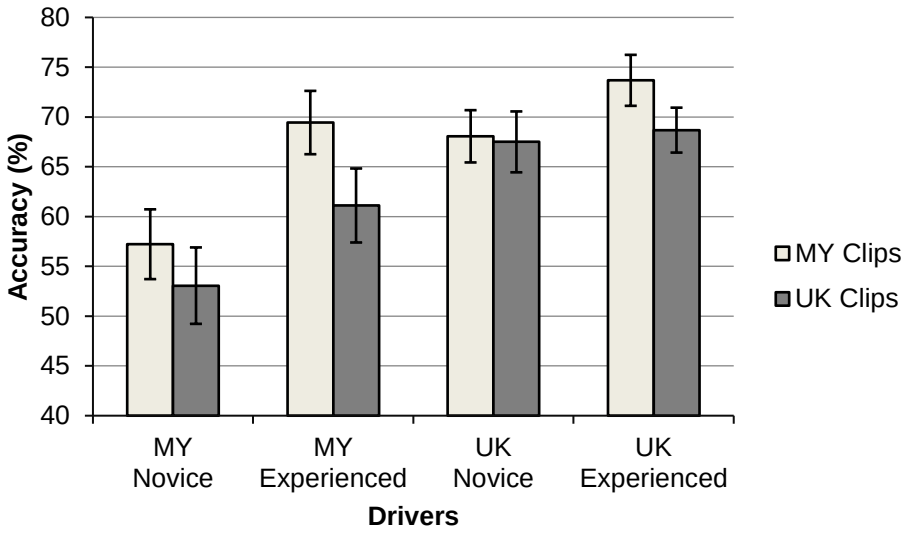

Figure 5-2: Accuracy for all drivers in Part II (videos and text options). Error bars represent standard error of the mean.

5.3.3 Prediction accuracy for videos

To analyze only prediction accuracy in the videos, a second ANOVA was conducted on Part II, i.e. only the video section. While this analysis was encompassed within the omnibus ANOVA above, this was done to allow a clearer direct comparison with the accuracy analysis run in Chapter 3 (Section 3.3.1.1). A 2 x 2 x 2 mixed design was used, analogous to the analysis run in Chapter 3. The between-groups factors were the origin country of the participant (driver origin: Malaysia or UK) and experience level (novice or experienced). The within-groups factor was the country where the clip was filmed (clip country: Malaysia or UK).

Mean scores for all groups of drivers are depicted above in **Figure 5-2**. All three possible main effects were found: experienced drivers outscored novices ($F_{1,69}$ = 7.07, p = .010, η^2_ρ = .093), UK drivers outscored Malaysian drivers ($F_{1,69}$ = 13.25, p = .001, η^2_ρ = .161), and finally, all participants were better at predicting hazards in

Malaysian videos ($F_{1,69} = 6.25$, $p = .015$, $\eta^2_\rho = .083$). There were no significant interactions (all *p*s > .19).

5.3.4 Answer distribution (Part I only)

A chi-square goodness of fit test was conducted for individual trials in Part I. Because Part I contained no incorrect answers in and of itself, the analysis was done across all four response options to determine whether all answers were equally likely to be chosen given no other context. An FDR-corrected α-value of .0338 was used to determine significance. **Table 5-2** reports results for Malaysian trials and **Table 5-3** reports results for UK trials. Out of 40 trials, response options were not equally distributed in 27 (13 Malaysian trials; 14 UK trials), suggesting that more often than not, participants felt one of the options was particularly likely or unlikely based on previous driving experience.

Table 5-2: Response distribution for Malaysian trials, Part I. Analysis conducted for all options, although Option 1 is the later, correct option in Part II.

Clip	Option 1	Option 2	Option 3	Option 4	χ^2	*p*
MY-M-01	11	18	15	29	9.79	.020*
MY-M-02	19	23	22	9	6.73	.081
MY-M-03	18	5	31	19	18.56	<.001*
MY-M-04	15	12	8	38	29.85	<.001*
MY-M-05	14	20	25	14	4.64	.200
MY-M-06	29	11	26	7	19.44	<.001*
MY-M-07	21	8	32	12	18.67	<.001*
MY-M-08	26	12	17	18	5.52	.137
MY-M-09	18	24	25	6	12.53	.006
MY-M-10	9	10	23	31	18.56	<.001*
MY-U-01	26	19	18	10	7.05	.070
MY-U-02	18	28	5	22	15.60	.001*
MY-U-03	28	13	13	19	8.26	.041
MY-U-04	21	4	15	33	24.04	<.001*
MY-U-05	27	18	21	7	11.55	.009*
MY-U-06	1	50	15	7	79.05	<.001*
MY-U-07	11	23	19	20	4.32	.229
MY-U-08	27	11	12	23	10.45	.015*
MY-U-09	19	10	27	17	8.04	.045
MY-U-10	10	25	11	27	13.30	.004*

* Significant at FDR-corrected $\alpha = .0338$

Table 5-3: Response distribution for UK trials, Part I. Analysis conducted for all options, although Option 1 is the later, correct option in Part II.

Clip	Option 1	Option 2	Option 3	Option 4	χ^2	*p*
UK-M-01	39	18	8	8	35.11	<.001*
UK-M-02	21	18	12	22	3.33	.344
UK-M-03	12	15	31	15	12.21	.007*
UK-M-04	16	24	19	14	3.11	.375
UK-M-05	14	22	18	19	1.79	.616
UK-M-06	28	7	16	22	13.19	.004*
UK-M-07	17	19	11	26	6.29	.098
UK-M-08	19	16	20	18	0.48	.923
UK-M-09	16	24	30	3	22.40	<.001*
UK-M-10	9	14	5	45	54.51	<.001*
UK-U-01	16	29	13	15	8.70	.034*
UK-U-02	5	36	15	17	27.55	<.001*
UK-U-03	16	27	23	7	12.64	.005*
UK-U-04	31	19	7	16	16.15	.001*
UK-U-05	31	8	14	20	15.82	.001*
UK-U-06	23	16	24	10	7.05	.070
UK-U-07	12	32	21	8	18.67	<.001*
UK-U-08	31	19	19	4	20.10	<.001*
UK-U-09	32	8	15	18	16.70	.001*
UK-U-10	25	11	8	29	17.47	.001*

* Significant at FDR-corrected $\alpha = .0338$

To investigate whether novice and experienced drivers differed in their prediction of what was likely to happen next, chi-square tests of independence were

conducted for individual trials in Part I, using driver experience and the four options for each clip as variables in a 2 x 4 contingency table for each trial. **Table 5-4** below shows the results.

Out of 40 trials, none survived an FDR correction, suggesting that for this experiment, there was no relation between a driver's experience level and likelihood of picking any particular traffic scenario.

Table 5-4: Tests of independence for whether experience was a factor in Part I predictions.

Item	Novice / Exp.	Option 1	Option 2	Option 3	Option 4	χ^2	*p*
MY-M-01	N	6	8	8	14	0.40	.940
	E	5	10	7	15		
MY-M-02	N	10	9	14	3	3.76	.288
	E	9	14	8	6		
MY-M-03	N	11	2	13	10	1.93	.586
	E	7	3	18	9		
MY-M-04	N	10	7	1	18	6.59	.086
	E	5	5	7	20		
MY-M-05	N	6	8	18	4	8.49	.037
	E	8	12	7	10		
MY-M-06	N	15	2	17	2	8.22	.042
	E	14	9	9	5		
MY-M-07	N	12	5	12	7	3.25	.355
	E	9	3	20	5		
MY-M-08	N	12	4	14	6	10.59	.014
	E	14	8	3	12		
MY-M-09	N	14	11	9	2	8.34	.040
	E	4	13	16	4		
MY-M-10	N	6	4	12	14	1.72	.632
	E	3	6	11	17		
MY-U-01	N	11	7	11	7	4.41	.221
	E	15	12	7	3		
MY-U-02	N	8	12	4	12	2.76	.430
	E	10	16	1	10		
MY-U-03	N	14	5	5	12	2.69	.442
	E	14	8	8	7		
MY-U-04	N	10	1	6	19	2.39	.495
	E	11	3	9	14		
MY-U-05	N	12	7	12	5	2.92	.404
	E	15	11	9	2		
MY-U-06	N	1	24	9	2	2.95	.399
	E	0	26	6	5		
MY-U-07	N	4	12	7	13	3.96	.265
	E	7	11	12	7		
MY-U-08	N	17	6	3	10	5.28	.152
	E	10	5	9	13		
MY-U-09	N	10	3	10	13	8.22	.042
	E	9	7	17	4		
MY-U-10	N	6	10	6	14	1.51	.679
	E	4	15	5	13		

Item	Novice / Exp.	Option 1	Option 2	Option 3	Option 4	χ^2	*p*
UK-M-01	N	19	10	2	5	2.73	.434
	E	20	8	6	3		
UK-M-02	N	7	8	8	13	4.60	.203
	E	14	10	4	9		
UK-M-03	N	4	10	17	5	4.94	.176
	E	8	5	14	10		
UK-M-04	N	9	13	8	6	1.16	.762
	E	7	11	11	8		
UK-M-05	N	9	14	8	5	7.25	.064
	E	5	8	10	14		
UK-M-06	N	13	4	9	10	0.70	.872
	E	15	3	7	12		
UK-M-07	N	8	9	4	15	1.53	.675
	E	9	10	7	11		
UK-M-08	N	12	4	10	10	5.53	.137
	E	7	12	10	8		
UK-M-09	N	8	15	11	2	3.95	.267
	E	8	9	19	1		
UK-M-10	N	4	7	3	22	0.32	.956
	E	5	7	2	23		
UK-U-01	N	10	16	4	6	3.82	.282
	E	6	13	9	9		
UK-U-02	N	1	20	9	6	4.30	.231
	E	4	16	6	11		
UK-U-03	N	7	15	11	3	0.76	.860
	E	9	12	12	4		
UK-U-04	N	15	8	2	11	4.03	.258
	E	16	11	5	5		
UK-U-05	N	17	3	7	9	0.98	.807
	E	14	5	7	11		
UK-U-06	N	9	7	13	7	3.09	.378
	E	14	9	11	3		
UK-U-07	N	7	14	11	4	0.87	.833
	E	5	18	10	4		
UK-U-08	N	15	10	9	2	0.12	.989
	E	16	9	10	2		
UK-U-09	N	21	2	8	5	8.74	.033
	E	11	6	7	13		
UK-U-10	N	17	5	1	13	8.13	.043
	E	8	6	7	16		

To investigate whether Malaysian and UK drivers differed in their prediction of what was likely to happen next, the above procedure was repeated using driver origin as a variable instead of experience. **Table 5-5** shows the results.

Out of 40 trials, none survived an FDR correction, suggesting that for this experiment, there was no relation between a driver's country of origin and likelihood of picking any particular traffic scenario.

Table 5-5: Tests of independence for whether driver origin was a factor in Part I predictions.

Item	Driver origin	Option 1	Option 2	Option 3	Option 4	χ^2	*p*
MY-M-01	MY	4	13	5	14	6.06	.109
	UK	7	5	10	15		
MY-M-02	MY	10	14	8	4	2.87	.411
	UK	9	9	14	5		
MY-M-03	MY	7	1	15	13	5.29	.152
	UK	11	4	16	6		
MY-M-04	MY	8	6	5	17	0.97	.807
	UK	7	6	3	21		
MY-M-05	MY	5	12	10	9	4.07	.254
	UK	9	8	15	5		
MY-M-06	MY	15	7	9	5	4.59	.205
	UK	14	4	17	2		
MY-M-07	MY	10	7	16	3	7.54	.057
	UK	11	1	16	9		
MY-M-08	MY	12	9	4	11	8.80	.032
	UK	14	3	13	7		
MY-M-09	MY	9	16	8	3	5.89	.117
	UK	9	8	17	3		
MY-M-10	MY	1	6	17	12	12.67	.005
	UK	8	4	6	19		
MY-U-01	MY	13	11	10	2	4.28	.232
	UK	13	8	8	8		
MY-U-02	MY	9	19	1	7	8.27	.041
	UK	9	9	4	15		
MY-U-03	MY	11	8	11	6	10.78	.013
	UK	17	5	2	13		
MY-U-04	MY	15	4	3	14	14.00	.003
	UK	6	0	12	19		
MY-U-05	MY	15	4	14	3	8.35	.039
	UK	12	14	7	4		
MY-U-06	MY	1	23	9	3	2.05	.562
	UK	0	27	6	4		
MY-U-07	MY	8	10	10	8	3.50	.320
	UK	3	13	9	12		
MY-U-08	MY	10	10	8	8	12.63	.006
	UK	17	1	4	15		
MY-U-09	MY	8	6	13	9	0.96	.812
	UK	11	4	14	8		
MY-U-10	MY	6	10	4	16	3.13	.372
	UK	4	15	7	11		

Item	Driver origin	Option 1	Option 2	Option 3	Option 4	χ^2	*p*
UK-M-01	MY	14	13	5	4	7.15	.067
	UK	25	5	3	4		
UK-M-02	MY	9	10	9	8	5.27	.153
	UK	12	8	3	14		
UK-M-03	MY	7	9	12	8	2.57	.463
	UK	5	6	19	7		
UK-M-04	MY	11	8	8	9	6.52	.089
	UK	5	16	11	5		
UK-M-05	MY	8	10	9	9	0.51	.917
	UK	6	12	9	10		
UK-M-06	MY	10	3	11	12	4.85	.183
	UK	18	4	5	10		
UK-M-07	MY	10	8	8	10	4.65	.199
	UK	7	11	3	16		
UK-M-08	MY	6	9	9	12	5.02	.171
	UK	13	7	11	6		
UK-M-09	MY	7	16	10	3	9.24	.026
	UK	9	8	20	0		
UK-M-10	MY	8	7	2	19	6.72	.081
	UK	1	7	3	26		
UK-U-01	MY	8	14	7	7	0.16	.983
	UK	8	15	6	8		
UK-U-02	MY	3	20	8	5	3.58	.310
	UK	2	16	7	12		
UK-U-03	MY	10	16	5	5	10.55	.014
	UK	6	11	18	2		
UK-U-04	MY	13	12	3	8	2.25	.522
	UK	18	7	4	8		
UK-U-05	MY	9	6	9	12	9.38	.025
	UK	22	2	5	8		
UK-U-06	MY	12	9	12	3	1.88	.598
	UK	11	7	12	7		
UK-U-07	MY	8	11	9	8	12.88	.005
	UK	4	21	12	0		
UK-U-08	MY	12	10	10	4	5.67	.129
	UK	19	9	9	0		
UK-U-09	MY	12	5	9	10	3.31	.346
	UK	20	3	6	8		
UK-U-10	MY	9	7	5	15	3.30	.348
	UK	16	4	3	14		

5.3.5 Distractor option plausibility (Part II only)

To determine whether the distractor options (i.e. the three incorrect options) were equally plausible, chi-square goodness of fit tests were performed on individual clips in Part II; the same analysis conducted in Chapter 3 (3.3.1.3). Ten clips were analyzed with an exact multinominal test due to having particularly low sample sizes (<5 expected responses in each cell), and a chi-square goodness of fit was conducted on the remaining 30. An FDR-corrected α-value of .0188 was used to determine significance. **Table 5-6** reports results for Malaysian trials and **Table 5-7** reports results for UK trials.

Out of 40 clips, participants' incorrect responses were not equally distributed in 15 (9 Malaysian clips; 6 UK clips), suggesting that for these clips, one or more of the distractor options was chosen substantially more often compared to the others. To ascertain any relationship between the distribution of the distractor options and how well a clip differentiated driving experience, effect sizes (given by Cohen's *d*) of the novice/experience difference were calculated for each clip. A Pearson's *r* correlation was then conducted using the chi-square value obtained above and its novice/experience effect size. The correlation was not significant ($r = -.096$, $p = .554$).

Table 5-6: Response distribution for Malaysian clips, Part II. χ^2 analysis conducted for distractor options only.

Clip	Correct response	Distractor 1	Distractor 2	Distractor 3	χ^2	*p*
MY-M-01	42	24	5	2	27.55	<.001*
MY-M-02	44	24	2	3	31.93	<.001*
MY-M-03	25	35	7	6	33.88	<.001*
MY-M-04	60	5	5	3	-	.831
MY-M-05	13	27	22	11	6.70	.035
MY-M-06	49	10	12	2	7.00	.030
MY-M-07	69	1	0	3	-	.333
MY-M-08	29	6	27	11	16.41	<.001*
MY-M-09	68	3	0	2	-	.383
MY-M-10	65	2	2	4	-	.744
MY-U-01	35	12	19	7	5.74	.057
MY-U-02	65	5	1	2	-	.296
MY-U-03	52	17	0	4	22.57	<.001*
MY-U-04	70	2	0	1	-	.778
MY-U-05	55	4	1	13	13.00	.002*
MY-U-06	45	19	9	0	19.36	<.001*
MY-U-07	61	3	7	2	-	.267
MY-U-08	57	5	7	4	0.88	.646
MY-U-09	48	1	20	4	25.04	<.001*
MY-U-10	28	32	8	5	29.20	<.001*

* Significant at FDR-corrected $\alpha = .0188$

Table 5-7: Response distribution for UK clips, Part II. χ^2 analysis conducted for distractor options only.

Clip	Correct response	Distractor 1	Distractor 2	Distractor 3	χ^2	*p*
UK-M-01	52	5	8	8	0.86	.651
UK-M-02	42	9	20	2	15.94	<.001*
UK-M-03	52	3	10	8	3.71	.156
UK-M-04	37	8	16	12	2.67	.264
UK-M-05	61	6	1	5	-	.178
UK-M-06	68	3	1	1	-	.630
UK-M-07	49	13	7	4	5.25	.072
UK-M-08	52	10	6	5	2.00	.368
UK-M-09	59	4	5	5	-	1
UK-M-10	46	18	6	3	14.00	.001*
UK-U-01	55	2	7	9	4.33	.115
UK-U-02	39	19	9	6	8.18	.017*
UK-U-03	25	11	18	19	2.38	.305
UK-U-04	35	27	7	4	24.68	<.001*
UK-U-05	40	12	16	5	5.64	.060
UK-U-06	51	21	1	0	38.27	<.001*
UK-U-07	48	10	8	7	0.56	.756
UK-U-08	32	4	1	36	55.07	<.001*
UK-U-09	26	19	18	10	3.11	.212
UK-U-10	45	9	8	11	0.50	.779

* Significant at FDR-corrected $\alpha = .0188$

5.4 Discussion

5.4.1 Hazard prediction with and without videos

There was very little overlap between the scenarios drivers selected in Part I, where they chose the most likely scenarios based on only text, and Part II, where they watched the associated videos before predicting the scenarios they thought would happen. We can safely conclude that participants were unable to guess the events that actually occurred by reading the scenarios alone, given that in Part I almost all driver groups selected the later, correct video scenarios at roughly chance levels. There does appear to be some consensus in the events thought most likely to happen, since participants' responses were unevenly distributed in 27 of 40 Part I trials; in other words, certain events were deemed particularly likely or unlikely in the majority of the Part I trials.

Driver experience and origin also have very little to do with the events drivers deem most likely to happen. There was no significant difference between driver groups, whether split by experience or origin, in choosing the later, correct video scenarios. Additionally, none of the chi-square tests conducted indicated any relation between drivers' level of experience and likelihood of picking any scenario, or drivers' country of origin and likelihood of picking any scenario. While these tests are generally underpowered, these findings do complement the Part I ANOVA results. Novice UK drivers did select correct UK events significantly above chance levels in Part I; however, it is difficult to conclude that they were generally better at predicting events than other driver groups, given the above chi-square results and the fact no group differences were highlighted in the omnibus ANOVA. Therefore, it seems likely that without contextual information, driving experience and culture have little to no effect on the events drivers deem most plausible.

5.4.2 Participants usually do not deem video scenarios the most likely to happen

While participants may agree somewhat on which events are most (and/or least) likely to happen on the road, these events are usually not the later, correct video scenarios. We can conclude that there is very little overlap between the two tasks, and given no other context, the correct scenario in the videos is usually not the one that participants deem most likely to happen.

However, this does not necessarily suggest that drivers are not choosing the events that are most likely to happen on the road; simply that their choices do not reflect the actual events that happen in the videos. It is possible that the scenarios represented in the videos are not necessarily the ones that are most likely to happen, since the clips selected for this experiment represented a wider variety of road situations and hazards than one might typically expect. The unmatched clips for instance were picked without restriction and thus included relatively unique hazards – particularly in Malaysia – such as a garbage bag falling off the back of a truck. The matched clips on the other hand represent roughly the same ten hazards in the UK and Malaysia – in other words half the total video hazards – again potentially skewing the proportion of hazards one might typically encounter. Additionally, because the hazards had to be salient to someone watching the videos, this adds a further element to the selection process; for instance, while a car braking ahead of a driver might be a common hazard on the road, it is often less salient in a video unless the braking is relatively abrupt. Conversely, vehicles coming from behind the driver are particularly common hazards in Malaysia, but are generally unsuitable for a hazard perception test because they have no precursor.

It is also possible that the scenarios depicted in the videos are indeed the ones most likely to happen, and participants simply did not choose these events with any consistency, although this seems less likely given the selection bias described above. However, the frequency of certain hazard types is certainly a factor to be considered when developing a test for licensing; the Queensland Transport Hazard Perception Test (QT-HPT) for instance largely focuses on hazards from the most frequent crash types involving young drivers in Queensland.

5.4.3 Specific task differences

Several interactions occurred in the analysis, mostly driven by the larger main effects occurring in Part II compared to Part I. We will discuss these three effects separately.

Firstly, drivers predicted Malaysian hazards more accurately than UK hazards in Part II, which is unsurprising given this was the case in the previous experiment using these videos (Chapter 3). However, drivers were slightly more likely to choose the correct UK scenarios in Part I. While this difference only trends towards significance, it was also consistent across all driver groups. Although some of these differences were negligible (notably, UK novices in Part II and Malaysian novices in Part I), all groups, without exception, had higher scores on the Malaysian videos compared to UK videos in Part II, and chose the correct UK video scenarios more often than Malaysian video scenarios in Part I. This pattern was fairly consistent regardless of drivers' country of origin, suggesting that the difference may stem from the different driving environments in general, and/or the particular scenarios chosen for this experiment. The higher unpredictability of the Malaysian driving environment and/or a self-selection bias for the Malaysian clips are likely contributors. More specifically, hazards are both more frequent and varied in

Malaysia compared to the UK, with more extreme examples occurring due to both higher hazard frequency and general driving environment. For instance, ten of the Malaysian hazards were selected without restriction, and include examples such as a garbage bag falling off the back of a truck, and pedestrians with luggage jaywalking in front of the camera car: events that are relatively unique for any driver, whether in Malaysia or the UK, and are unlikely to be chosen as the most plausible. In fact, very few participants chose these particular scenarios in Part I; 10 participants out of 73 selected the garbage bag option, and only 1 out of 73 selected the pedestrian option. These differences are large enough that they could alone could be driving the Part I effect, although as mentioned the general driving environment in Malaysia is likely to contribute as well.

Secondly, experienced drivers outscored novices in Part II, another unsurprising finding consistent with the previous experiments. However, there was no difference between these groups in Part I, where both groups chose the correct video scenario equally often, and roughly at chance levels. While novice UK drivers chose UK events above chance levels, as discussed above, it is difficult to conclude that they were significantly better than other driver groups given that no group differences emerged in the omnibus ANOVA nor chi-square tests. As discussed earlier, the correct scenario in the videos is usually not the one that participants think most likely to happen. It therefore seems likely that neither group of drivers was able to guess the correct video scenarios to any significant degree.

Lastly, UK drivers significantly outscored Malaysian drivers in Part II, another finding consistent with Chapter 3. However, there was no difference between these groups in Part I, where, again, both groups chose the correct video scenario equally often and roughly at chance levels (with the exception of UK novice drivers and UK

events, as described above). As with the above interaction, the correct scenario in the videos is usually not the one that participants think most likely to happen, and this particular interaction is primarily driven by the main effect found in Part II.

5.4.4 Implications

As discussed, the correct scenario in the videos is usually not the one participants deem most likely to happen. This suggests that prior knowledge of road events plays a minor role in hazard perception performance, at least as measured by the "What Happens Next?" test. This finding is perhaps unsurprising, given that the results of Chapter 3 indicated that familiarity with a country's hazards was not a primary driver of performance compared to the ability to extrapolate events from video scenarios, regardless of familiarity. If this were the case, we would have expected some cultural differences in Chapter 3; for instance drivers should have been better at predicting hazards in their country of origin. However there were no cultural interactions; in fact, UK drivers outscored Malaysian drivers, and all drivers were better able to predict Malaysian hazards compared to UK hazards, suggesting that the "What Happens Next?" test in a multiple choice format is reasonably independent of culture.

The results indicate that neither driving experience nor culture greatly influence which hazards drivers deem most likely to happen. However, when data from all drivers was pooled, there did appear to be some consensus among all drivers about which events are likely to happen, suggesting drivers in general may have similar ideas about hazard frequency, but driving experience and/or driving environment do not play a major part in this. On the other hand, data from Chapter 3 indicated that both driving experience and to a lesser extent environment affects drivers' hazard perception skill. Combined with the above observations, we can

conclude that without viewing actual road scenarios, drivers generally agree on which events are likely to happen regardless of environment or experience; however, once drivers view road scenarios, experience in particular plays a much larger role in their ability to anticipate events.

From a practical perspective, it appears that a questionnaire-only hazard perception test, at least one based on this particular paradigm, has limited utility. However, other questionnaire-based tasks may still be viable. For instance, one might argue that this particular task is an unsuitable measure of hazard perception; since crashes are relatively rare events, being aware of infrequently occurring hazards may be as important, or even more so, than being aware of frequently occurring hazards. An alternative task could be one where participants predict all hazards regardless of plausibility and discount impossible ones, rather than picking the hazards that seems most likely.

CHAPTER 6

CAN REDUCED RESOLUTION VIDEOS EFFECTIVELY MEASURE HAZARD PERCEPTION ABILITY?

Abstract

The hazard perception skill of a driver refers to the ability to identify potentially dangerous situations on the road. It is typically tested by presenting drivers with videos filmed on the road, and asking them to press a response button as soon as they detect developing hazards. In the UK and Australia, drivers must pass a hazard perception test as part of the licensing procedure, and research has suggested hazard perception testing may also be effective in lower-income countries with higher crash rates such as Malaysia. While almost all hazard perception research has been conducted in person and with high definition videos, it is not always feasible to replicate laboratory conditions; for instance, in Queensland, Australia, an online hazard perception test is used for driver licensing, which requires much lower quality videos. In the present study, drivers viewed clips of driving scenes that had been visually degraded to different degrees and predicted events that were likely to happen after the clips had been occluded, a hazard perception task called the "What Happens Next?" test. Similar to previous versions of this test, experienced drivers outperformed novices, suggesting that videos of greatly reduced quality can still differentiate driving experience.

6.1 Introduction

In the three previous chapters, we found that different versions of the "What Happens Next?" test differentiated experience in Malaysian and UK drivers. Chapters 3 and 5 differentiated experience in both Malaysian and UK drivers using Malaysian (Chapter 3) and UK (both Chapters 3 and 5) videos, and Chapter 4 differentiated experience in Malaysian drivers using Malaysian videos. These results are promising for hazard perception testing in Malaysia, particularly Chapters 3 and 5, as these tests used a multiple choice response format which allows for wide-scale deployment. Furthermore, in Chapter 5 both Malaysian and UK videos differentiated experience for Malaysian and UK drivers, suggesting that the current version of the test may be suitable for use in both these countries.

As with almost all hazard perception research, all experiments thus far have been administered under controlled conditions, usually with a researcher present during test-taking. While these are ideal conditions for research purposes, it is not always possible to conduct large-scale tests under similar conditions. For instance, in Queensland, Australia, 10.5% of the population live in rural and remote areas, where residents may have difficulty visiting a driving center in person (Queensland Treasury and Trade, 2012). The Queensland Transport Hazard Perception Test (QT-HPT) can therefore be taken online, on an applicant's own computer. Similar issues may arise in developing countries where rural populations tend to be high; for example, 26% of Malaysia's population resides in rural areas, compared to Queensland's 10.5% (Toroyan, 2013).

While a general lack of control in online testing is certainly cause for concern, another issue this creates is the quality of the videos used in hazard perception tests. The QT-HPT requires a 30 – 60 MB download: for a 15-item test, this averages a

maximum of 4 MB per video, far smaller than most videos used in hazard perception research, which are usually presented in high definition. As a comparison, the 40 videos used in Chapter 3 averaged 25.7 MB a clip, and were presented at a resolution of 1280 x 720. While it is certainly possible to achieve some degree of compression without a significant loss in video quality, compression to an acceptable level for online testing would likely require a noticeable drop in quality. Furthermore, internet access in Malaysia and other developing countries may be more limited in rural areas compared to Queensland, possibly necessitating an even greater degree of compression.

The present chapter therefore explores the effectiveness of the "What Happens Next?" test when videos of much lower quality are used. The visual quality of the clips used in Chapters 3 and 5 was degraded to two different degrees (medium quality and low quality), and presented to experienced and novice Malaysian drivers in a mixed design. While it seems highly likely that overall accuracy scores will decrease with lower quality videos, degrading the videos may also affect novice and experienced drivers differently. The first possibility is that degraded videos make the task particularly difficult for novice drivers, in which case scores should decrease for all drivers but especially for novices, widening the gap between driver groups. Alternatively, degrading clips may obscure some of the cues that experienced drivers use to make predictions, which would remove some of the advantages that they would normally be able to exploit; this would therefore result in scores decreasing for all drivers but especially for experienced drivers. The final possibility is that degradation makes prediction more difficult but affects all drivers equally, in which case overall scores should decrease but the gap between the novice and experienced driver groups should remain fairly constant.

6.2 Methods

6.2.1 Participants

Thirty-seven participants were recruited from Malaysia, all of whom held full or provisional driving licenses. Participants were split into two groups based on their driving experience, resulting in 19 novice drivers (mean age of 17.9 years and licensing time of 4.7 months) and 18 experienced drivers (mean age of 28.2 years and licensing time of 127.1 months). All participants received monetary compensation. None of the participants in this study had participated in any of the previous experiments.

6.2.2 Stimuli and apparatus

The original stimuli were the same videos and their corresponding text options used in Chapter 3, consisting of 20 clips from Malaysia and 20 from the UK, as well as two practice clips from each country. In Chapters 3 and 5, all clips were presented at a resolution of 1280 x 720, which we will refer to as *high quality*. The videos were subsequently edited to reduce the amount of visual information present in the clips. This was done by reducing the original clips to a considerably lower resolution of 320 x 180 (*low quality* condition) and 480 x 270 (*medium quality* condition), and then re-saving these low resolution clips at a final resolution of 1280 x 720 for use in the experiment. Examples can be seen below in **Figure 6-1** and **Figure 6-2**. The clips were presented using PsychoPy 1.76 on a 13" laptop with the screen set to maximum brightness .

Figure 6-1: Comparison stills from a video filmed in Malaysia. From top to bottom: low quality, medium quality, high quality.

Figure 6-2: Comparison stills from a video filmed in the UK. From top to bottom: low quality, medium quality, high quality.

6.2.3 Design

A 2 x 2 x 2 mixed design was used. The between-groups factor was the experience level of the drivers (novice or experienced) and the within-groups factors were the country where the clip was filmed (clip country: Malaysia or UK), and the quality of the clip (clip quality: low or medium). From the original set of 20, the Malaysian and UK clips were split into two further sub-groups of 10 clips each, resulting in four total sub-groups: Malaysia Group A, Malaysia Group B, UK Group A, and UK Group B. Both sub-groups for each country were equated for accuracy and standard deviation for both Malaysian and UK drivers, based on the scores obtained in Chapter 3 (all *ps* > .94); in other words, the mean score and standard deviation for UK drivers on Malaysia Group A clips (for instance) was as close as possible to the mean score and standard deviation for UK drivers on Malaysia Group B clips. Mean scores and standard deviations for all clip sub-groups are reported in **Table 6-1**. Each participant viewed either all Group A clips at low quality and Group B clips at high quality, or vice versa; this was counterbalanced across participants. All participants also viewed four practice clips, one from each possible sub-group, using the same counterbalancing as above.

The stimuli were presented in two blocks of 20 clips, using mini-blocks of four that each contained one clip from every possible sub-group. No two videos from the same sub-group were shown directly after each other. Within these restrictions, the order of presentation was randomly generated for each participant, as were the corresponding multiple choice options shown after the video ended.

Table 6-1: Mean scores (%) and standard deviations for each clip sub-group, based on participants' scores in Chapter 3.

	Malaysian clips		UK clips	
	Group A	**Group B**	**Group A**	**Group B**
Malaysian drivers	66.2 (23.7)	66.2 (23.8)	57.6 (21.3)	57.6 (20.7)
UK drivers	71.0 (24.8)	70.8 (24.9)	65.3 (19.7)	65.3 (20.3)

6.2.4 Procedure

After giving informed consent, participants completed a brief demographic questionnaire and were then seated in front of the laptop. Participants were given the same instructions as in Chapter 3. Participants were informed that each clip contained a driving scenario leading up to a hazardous event, however the clips would end immediately before this event actually occurred and their task was to predict what the event was by selecting the correct scenario out of four possible options. They were informed that in every case, one and only one of the four scenarios had actually taken place and there was therefore a correct answer for each clip. It was also emphasized that their task was not to choose the event that they felt was the most hazardous, but the one that was most likely to have occurred.

Before starting the main experiment, participants attempted four practice clips, two from Malaysia and two from the UK. These practice clips were not used in the subsequent experiment. After the practice clips they were able to ask questions or seek clarification. Participants were not given any feedback as to the correct scenarios at any point during the practice clips or main experiment.

A short line of text was displayed for 1 second before each clip, indicating participants' progress through the block. After watching each clip, participants

selected the scenario they thought most likely to occur by pressing the corresponding number on the laptop keyboard (1, 2, 3, or 4). After selecting their answer, the progress text appeared to signal the beginning of the next clip (or end of the block, if appropriate), and the process was repeated until the end of the block. After the first block, participants were given the opportunity to take a brief break, and the process was repeated.

6.3 Results

6.3.1 Accuracy

Accuracy scores are summarized in **Figure 6-3**. To confirm that splitting the clips into sub-groups had not affected performance, two independent t-tests were conducted for the Malaysian and UK clips comparing both counterbalanced participant groups; in other words, participants who had viewed high quality Group A clips and low quality Group B clips, and those who had viewed low quality Group A clips and high quality Group B clips (both *p*s > .55). Accuracy scores were then analyzed using a 2 x 2 x 2 mixed ANOVA. The between-groups factor was the experience level of the drivers (novice or experienced) and the within-groups factors were the country where the clip was filmed (clip country: Malaysia or UK), and the quality of the clip (clip quality: low or high). Two main effects were found: experienced drivers outscored novices ($F_{1,35} = 7.60$, $p = .009$, $\eta^2_\rho = .178$), and all drivers had higher scores on Malaysian clips compared to UK clips ($F_{1,35} = 13.88$, $p = .001$, $\eta^2_\rho = .284$). There were no significant interactions.

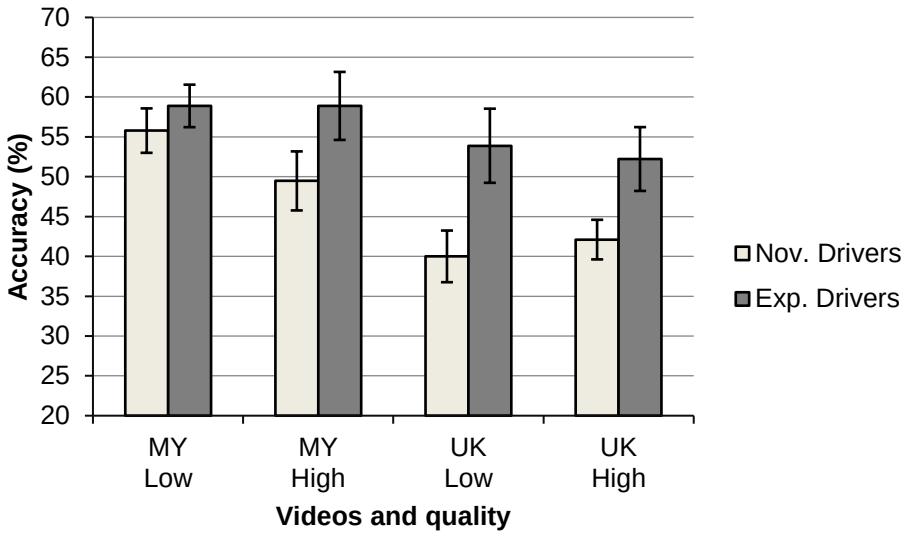

Figure 6-3: Accuracy scores for all drivers across video conditions. Error bars represent standard error of the mean.

6.3.2 Distractor option plausibility

To determine whether the distractor options (i.e. the three incorrect options) were equally plausible, goodness of fit tests were performed on individual clips. Only incorrect options chosen by participants were included in this analysis; correct responses were excluded. Thirteen clips were analyzed with an exact multinominal test due to having particularly low sample sizes (<5 expected responses in each cell), and a chi-square goodness of fit was conducted on the remaining 27. An FDR-corrected α-value of .02 was used to determine significance. Results are reported in **Table 6-2** and **Table 6-3**.

Out of 40 clips, participants' incorrect answers were not equally distributed in 16 (10 Malaysian clips; 6 UK clips), suggesting that for these clips, one or more of the distractor options was chosen substantially more often compared to the others. To ascertain any relationship between the distribution of the distractor options and how well a clip differentiated driving experience, effect sizes (given by Cohen's *d*) of the novice/experience difference were calculated for each clip. A Pearson's *r* correlation

was then conducted using the chi-square value obtained above and its novice/experience effect size. The correlation was not significant (r = -.188, p = .246).

Table 6-2: Response distribution for Malaysian clips. χ^2 analysis conducted for only distractor options.

Clip	Correct response	Distractor 1	Distractor 2	Distractor 3	χ^2	*p*
MY-M-01	11	18	15	29	9.79	.020*
MY-M-02	19	23	22	9	6.73	.081
MY-M-03	18	5	31	19	18.56	<.001*
MY-M-04	15	12	8	38	29.85	<.001*
MY-M-05	14	20	25	14	4.64	.200
MY-M-06	29	11	26	7	19.44	<.001*
MY-M-07	21	8	32	12	18.67	<.001*
MY-M-08	26	12	17	18	5.52	.137
MY-M-09	18	24	25	6	12.53	.006
MY-M-10	9	10	23	31	18.56	<.001*
MY-U-01	26	19	18	10	7.05	.070
MY-U-02	18	28	5	22	15.60	.001*
MY-U-03	28	13	13	19	8.26	.041
MY-U-04	21	4	15	33	24.04	<.001*
MY-U-05	27	18	21	7	11.55	.009*
MY-U-06	1	50	15	7	79.05	<.001*
MY-U-07	11	23	19	20	4.32	.229
MY-U-08	27	11	12	23	10.45	.015*
MY-U-09	19	10	27	17	8.04	.045
MY-U-10	10	25	11	27	13.30	.004*

* Significant at FDR-corrected $\alpha = .02$

Table 6-3: Response distribution for UK clips. χ^2 analysis conducted for only distractor options.

Clip	Correct response	Distractor 1	Distractor 2	Distractor 3	χ^2	*p*
UK-M-01	56	4	5	3	-	.935
UK-M-02	47	1	14	14	11.66	.003*
UK-M-03	52	3	16	5	12.25	.002*
UK-M-04	49	7	9	11	0.89	.641
UK-M-05	65	4	4	4	-	1
UK-M-06	66	7	0	3	-	.022
UK-M-07	49	11	5	11	2.67	.264
UK-M-08	38	18	6	15	6.00	.050
UK-M-09	16	32	16	12	11.20	.004*
UK-M-10	53	6	13	5	4.75	.093
UK-U-01	61	5	10	1	7.63	.022*
UK-U-02	39	11	3	24	17.74	<.001*
UK-U-03	23	19	15	20	0.78	.678
UK-U-04	36	23	16	1	18.95	<.001*
UK-U-05	50	13	3	10	6.08	.048
UK-U-06	62	3	12	0	15.60	<.001*
UK-U-07	37	17	10	13	1.85	.397
UK-U-08	60	10	3	4	5.06	.080
UK-U-09	34	7	15	21	6.88	.032
UK-U-10	47	9	9	11	0.28	.871

* Significant at FDR-corrected $\alpha = .02$

6.4 Discussion

The first section of this discussion focuses on the present results but in the context of both the present and previous studies, given that this is the third study in this thesis that uses the same videos and task. The second section specifically compares and summarizes the results of all three studies.

The first section of this discussion focuses on the present results but in the context of both the present and previous studies, given that this is the third study in this thesis that uses the same videos and task. The second section specifically compares and summarizes the results of all three studies using the same videos (Chapters 3, 5, and 6).

6.4.1 Degradation effect

6.4.1.1 *Present chapter only*

Accuracy analyses showed no effect of the degradation manipulation; videos presented at both low and medium quality differentiated driving experience equally well, and there was no main effect of degradation. This suggests that the information lost from medium to low levels of degradation was equally important – or rather, unimportant – for all participants, regardless of their driving experience.

From a practical perspective, this corroborates the results reported by Horswill, Hill, et al. (2015), which linked drivers' reaction times on the Queensland Transport Hazard Perception Test (QT-HPT) to their later crash rates. The QT-HPT is conducted online, and thus uses videos that are significantly reduced in quality; however, Horswill, Hill, et al.'s results suggest this does not affect its validity as a measure of hazard perception. Similarly, the present study confirmed that "What Happens Next?" test retains its effectiveness as a measure of hazard perception, even when videos are significantly reduced in quality. This is especially promising

because the "What Happens Next?" test and the QT-HPT use different paradigms to measure hazard perception, suggesting that future hazard perception tests using either paradigm can be conducted online and remain effective.

6.4.1.2 *Combined chapter results*

The results of the present chapter alone suggest the degradation manipulation had no effect, when comparing low and medium quality videos. However, if we also consider the results of Chapters 3 and 5, where the same videos were used but in high quality, this suggests that reducing video quality may have some effects after all. First and most obviously, overall accuracy scores are much lower in the present study than in previous chapters (51.2%, compared to 61.9% and 60.1% for Malaysian drivers in Chapters 3 and 5 respectively), which presumably reflects the increased difficulty of the task when using reduced quality videos. More interestingly, the experienced/novice group differences in the present study are considerably larger than either of the previous experiments ($\eta^2_\rho = .178$, compared to $\eta^2_\rho = .057$ and $\eta^2_\rho = .093$ in Chapters 3 and 5 respectively). However, this could also be explained by greater levels of driving experience in the present study, as the effect sizes above also correspond with mean licensing time for the experienced driver group (127.1 months in the present study; 55.2 and 87.2 months in Chapters 3 and 5 respectively). The greater experience differentiation in the present chapter could therefore be due to either reduced quality videos, greater driver experience, or more likely, some combination thereof; while it is not possible to separate these effects with the data currently available, this presents an obvious avenue for future work.

Given the above observations, it seems likely that the effects of degrading video quality also depend on the level of degradation that is applied. It also appears that the reduction in video quality, from high to medium/low, removed information

that is necessary for hazard perception (while the reduction from medium to low quality videos did not, as discussed in Section 6.4.1.1), as accuracy scores in the present chapter are considerably lower than in previous chapters. At present, it is unclear whether the information that was removed by the reduction in video quality was being used by both novice and experienced drivers to an equal extent.

Finally, we should note that the terms high/medium/low have been applied largely for ease of use, and represent ordinal rather than interval measurements: the high quality videos presented in Chapters 3 and 5 had original resolutions of 1280 x 720, while the medium and low quality videos used in the present experiment had reduced resolutions of 480 x 270 and 320 x 180 respectively (although all videos were presented at 1280 x 720; see Section 6.2.2 for details). Consequently, the drop in quality from high to medium/low videos is fairly noticeable, while the difference between the medium and low quality videos is much subtler; see **Figure 6-1** and **Figure 6-2** for visual examples. We might therefore expect greater differences between high and medium/low quality videos – i.e. Chapters 3 and 5 vs. the present chapter –, compared to medium and low quality videos – i.e. the experiment in the present chapter. This does indeed seem to be the case, as discussed above.

6.4.2 Overall comparison to previous chapters

The main effects in the present study are entirely consistent with the two previous experiments using these videos; experienced drivers outscored novices, and scores were higher in the Malaysian set of videos than in the UK set. Interestingly, similar to Chapter 5, there was again no interaction of experience and clip country, and the Malaysian and UK videos differentiated experience equally well, as opposed to Chapter 3 where only the UK videos differentiated experience. This may be due to the larger experience gap between driver groups in the latter two experiments; it is

possible that the UK videos may be more sensitive overall, but given enough experience the Malaysian videos are also capable of differentiating driver groups. This corresponds with several studies that have reported differences between experienced drivers and 'expert' drivers such as police drivers or driving instructors (Crundall et al., 2012, 2003; Groff & Chaparro, 2003; Underwood et al., 2002), suggesting that task performance does improve with greater levels of experience. Previous research indirectly supports this idea, where generally, differences have been found with larger experience gaps between groups (Horswill et al., 2008; Wallis & Horswill, 2007; Wetton et al., 2010), and several studies using a relatively small experience gap have failed to find a group difference (Chapman & Underwood, 1998; Crundall et al., 2002; Underwood et al., 2013), although there have been some exceptions where no differences were found with even a highly experienced group (Crundall et al., 2003; Sagberg & Bjørnskau, 2006).

Given the present results and those of Chapters 3 and 5, we can tentatively conclude that both video quality and level of experience affect this test's experience differentiation. We can infer this from comparing the effects of experience in all three studies; the effect is smallest in Chapter 3 (55.2 months mean licensing time for experienced group; normal video quality; $\eta^2_\rho = .057$), slightly higher in Chapter 5 (87.2 months; normal video quality; $\eta^2_\rho = .093$), and by far the largest in the present chapter (127.1 months; lower video quality; $\eta^2_\rho = .178$). (Note that in all three studies, the mean novice licensing time was under a year.) As discussed above, it seems possible that the degradation manipulation also influenced these results, but overall, given the present results, it is difficult to come to a more specific conclusion than both these factors likely affect test differentiation. It is also possible that the additional Part I task in Chapter 5 was another factor, and reading the various

scenarios before watching the corresponding videos affected drivers' responses; however, this seems unlikely as most driver groups had similar scores in each clip group in Chapters 3 and 5, i.e. both chapters with normal quality clips.

Distractor response distribution is also similar between experiments; 25 of 40 clips have similar response distribution (i.e. one of the distractor options was particularly likely or unlikely) in all three studies (Chapters 3, 5, and 6), and 30 of 40 have similar distribution between Chapters 3 and 5, with normal quality clips. Given the present and previous results, it seems unlikely that response distribution among distractors has a major influence on a clip's ability to differentiate experience. It may be the case that as long as some of the distractors are reasonably plausible, this is sufficient to create a good test item.

The distractor and accuracy results both suggest that clips are viewed relatively consistently and retain their individual characteristics regardless of slight differences in presentation. This implies that, as discussed in previous chapters, although various factors such as visual information can affect experience differentiation, the individual properties of a clip contribute largely to its success in differentiating experience.

Overall, the results are encouraging in their general consistency, especially given that all three experiments using these videos have differed slightly in presentation. The next and final empirical chapter collapses data from Chapters 3, 5, and the present one to analyze the effectiveness of the present "What Happens Next?" test across all participants.

CHAPTER 7

PSYCHOMETRIC PROPERTIES OF THE "WHAT HAPPENS NEXT?" TEST USED IN THIS THESIS

Abstract

An item analysis was conducted using data from three "What Happens Next?" experiments that had used the same videos (Chapters 3, 5, and 6). Reliability and validity were analyzed, as well as how effectively individual test items differentiated driving experience. After selecting the clips that best discriminated experience and overall performance, the original item pool of 40 was narrowed to 15. The final group of 15 clips showed a large effect size for experience differentiation, but internal reliability was somewhat low. While not meeting the standards for a high-stakes test such as licensing, the results seem promising for future hazard perception test development using the "What Happens Next?" paradigm.

7.1 Introduction & Methods

While the experiments described in this thesis have largely utilized the "What Happens Next?" test as an experimental paradigm, one of the key practical issues is its viability for use in driver licensing. Since a large number of drivers participated across the three experiments described in Chapters 3, 5, and 6, this provided the opportunity to evaluate the performance of the test used in these experiments as a potential driver licensing tool.

In order to gauge consistency of the results across all three experiments, measures from Chapters 3, 5, and 6 were first correlated. Following this, an item analysis was conducted; participant data across the three experiments was collapsed and the test was analyzed as a four-item multiple choice test, with all 40 videos (or test items) included. As in a standard item analysis, reliability and validity were analyzed, as well as individual items to gauge their contribution to test effectiveness. Performance for novice and experienced drivers is also reported for each item, collapsed across all experiments. For all Chapter 5 analyses, only Part II results were used since Part I did not incorporate videos.

7.2 Results

187 drivers in total participated across the three experiments, with no repeat participants. The total number of items was 40 and each item had four options in total. In Chapters 5 and 6, all participants answered all items; in Chapter 3, there were 16 invalid responses due to incorrect keypresses (0.5% of the total responses for that experiment and 0.2% for all three experiments combined). These were deemed incorrect for all analyses except those involving distractor options, where they were treated as non-answers and excluded from analysis.

Section 7.2.1 reports consistency for individual videos by comparing the following three measures across all three experiments: how accurately participants predicted hazards (measured by mean accuracy), how well videos differentiated experience (measured by effect size), and the plausibility of the three distractor options for each video (measured by chi-square value).

Section 7.2.2 collapses participant data across all three experiments and reports the item analysis. Sections 7.2.2.2 and 7.2.2.3 report test-wide statistics such as mean scores and reliability and validity measures, and the remaining sections

report measures for individual items. The following measures are included for each item: Item Difficulty, biserial correlation coefficient (r_{bis}), mean scores for both novice and experienced drivers, effect size of the novice/experienced driver score difference, and answer frequency for each item option (with the first option always being correct). These measures are explained in more detail in their individual sections. All effect sizes of the novice/experience difference are reported as Cohen's *d*; positive *d*-values indicate that experienced drivers had higher scores for that item, and negative *d*-values indicate that novices had higher scores.

Finally, 7.2.3 summarizes the results of eliminating less useful test items, which was done on the basis of how effectively clips differentiated experience and general test performance.

7.2.1 Consistency across experiments

Data from all three analyses is summarized in **Table 7-1** and **Table 7-2**. Pearson's *r* correlations were conducted for all three measures.

7.2.1.1 *Accuracy*

Mean scores for each individual video for all drivers were compared across all three experiments. All experiments showed strong positive correlations (all *p*s < .001): for Chapters 3 and 5, $r = .749$; Chapters 3 and 6, $r = .818$; Chapters 5 and 6, $r = .742$.

7.2.1.2 *Experience differentiation*

Experience differentiation was measured by the effect size of the novice/experience performance gap. None of the correlations were significant. Chapter 3's experiment showed a non-significant positive correlation with both

Chapter 5 and 6 (r = .094, p = .563; r = .196, p = .225 respectively), and Chapters 5 and 6 showed a non-significant negative correlation (r = -.076, p = .640).

7.2.1.3 *Distractor plausibility*

Distractor plausibility was measured by the goodness of fit chi-square value that was obtained by comparing the three distractor options (see Sections 3.3.1.3, 5.3.5, and 6.3.2 for further detail). All experiments showed strong positive correlations: for Chapters 3 and 5, r = .546, p < .001; Chapters 3 and 6, r = .477, p = .002; Chapters 5 and 6, r = .452, p = .003.

Table 7-1: Various measures for individual Malaysian clips in each experiment from Chapters 3, 5, and 6.

Item / Clip	Accuracy (%)			Effect size (Cohen's *d*)			χ^2 for distractor options		
	Chap. 3	Chap. 5	Chap. 6	Chap. 3	Chap. 5	Chap. 6	Chap. 3	Chap. 5	Chap. 6
MY-M-01	61.0	57.5	48.6	-0.02	0.37	0.26	18.20	27.55	14.63
MY-M-02	72.7	60.3	62.2	0.65	0.61	-0.04	12.10	31.93	-
MY-M-03	29.9	34.2	18.9	-0.38	0.08	0.16	50.30	33.88	15.20
MY-M-04	93.5	82.2	75.7	-0.11	0.06	0.34	-	-	-
MY-M-05	32.5	17.8	29.7	-0.15	0.68	-0.57	7.41	6.70	0.08
MY-M-06	59.7	67.1	70.3	0.58	0.33	-0.39	8.58	7.00	-
MY-M-07	97.4	94.5	89.2	0.32	0.23	0.71	-	-	-
MY-M-08	53.2	39.7	29.7	-0.34	-0.03	-0.08	11.37	16.41	10.23
MY-M-09	85.7	93.2	78.4	0.06	0.10	0.50	-	-	-
MY-M-10	77.9	89.0	51.4	-0.20	0.16	0.61	6.12	-	2.33
MY-U-01	29.9	47.9	18.9	-0.04	-0.03	-0.11	14.78	5.74	15.00
MY-U-02	92.2	89.0	86.5	-0.20	0.34	0.13	-	-	-
MY-U-03	70.1	71.2	51.4	0.38	0.16	0.38	22.52	22.57	21.00
MY-U-04	96.1	95.9	91.9	-0.14	-0.14	0.60	-	-	-
MY-U-05	74.0	75.3	54.1	0.10	0.24	0.06	7.30	13.00	7.18
MY-U-06	62.3	61.6	54.1	0.03	0.32	-0.15	29.45	19.36	19.18
MY-U-07	93.5	83.6	78.4	-0.11	0.13	-0.29	-	-	-
MY-U-08	84.4	78.1	37.8	0.13	0.39	0.74	-	0.88	4.26
MY-U-09	68.8	65.8	62.2	0.21	0.04	0.40	5.25	25.04	-
MY-U-10	37.7	38.4	24.3	-0.36	0.25	-0.09	52.26	29.20	6.50

Table 7-2: Various measures for individual UK clips in each experiment from Chapters 3, 5, and 6.

Item / Clip	Accuracy (%)			Effect size (Cohen's *d*)			χ^2 for distractor options		
	Chap. 3	Chap. 5	Chap. 6	Chap. 3	Chap. 5	Chap. 6	Chap. 3	Chap. 5	Chap. 6
UK-M-01	84.4	71.2	70.3	0.13	0.41	0.32	-	0.86	-
UK-M-02	61.0	57.5	62.2	0.64	0.03	-0.04	11.66	15.94	-
UK-M-03	67.5	71.2	67.6	0.61	0.16	0.68	12.25	3.71	-
UK-M-04	63.6	50.7	43.2	0.09	0.19	-0.39	0.89	2.67	1.14
UK-M-05	84.4	83.6	78.4	0.58	0.44	0.81	-	-	-
UK-M-06	85.7	93.2	81.1	0.06	-0.11	-0.16	-	-	-
UK-M-07	63.6	67.1	43.2	0.42	0.21	0.26	2.67	5.25	13.71
UK-M-08	49.4	71.2	62.2	-0.08	0.82	0.40	6.00	2.00	-
UK-M-09	20.8	80.8	24.3	0.14	-0.01	0.68	11.20	-	4.79
UK-M-10	68.8	63.0	43.2	0.21	-0.08	0.26	4.75	14.00	1.14
UK-U-01	79.2	75.3	43.2	-0.01	-0.02	0.26	7.63	4.33	3.43
UK-U-02	50.6	53.4	21.6	0.39	-0.03	0.29	17.74	8.18	28.41
UK-U-03	29.9	34.2	35.1	0.19	0.19	0.38	0.78	2.38	1.00
UK-U-04	46.8	47.9	16.2	0.34	-0.25	0.31	18.95	24.68	8.19
UK-U-05	64.9	54.8	37.8	0.14	0.48	0.26	6.08	5.64	1.65
UK-U-06	81.8	69.9	59.5	-0.01	0.10	-0.37	15.60	38.27	14.80
UK-U-07	48.1	65.8	29.7	0.18	-0.19	0.39	1.85	0.56	4.92
UK-U-08	77.9	43.8	48.6	-0.20	-0.09	0.26	5.06	55.07	7.68
UK-U-09	44.2	35.6	35.1	0.02	-0.21	0.15	6.88	3.11	0.75
UK-U-10	61.0	61.6	35.1	0.08	0.32	0.62	0.28	0.50	1.75

7.2.2 Item analysis

7.2.2.1 *Overall test statistics*

The mean score across all participants was 24.9, or 62.3%, and the median score was 26 or 65%. Minimum and maximum scores for participants were 12 (30%) and 34 (85%) respectively, while for test items, scores ranged from 26.2% to 95.2%. The standard deviation of the entire test was 4.96, or 12.4%.

7.2.2.2 *Reliability and validity*

Cronbach's alpha was used to measure internal consistency; because this test used dichotomous items, this gives the same result as the Kuder-Richardson Formula 20 (KR-20). Across all participants, $\alpha = .69$. Validity was not measured directly, since convergent validity with driving experience is implicit in all previous accuracy analyses using these items (Chapters 3, 5 and 6). There has been a significant performance difference between novice and experienced drivers in all instances of this test's administration, suggesting that the test as a whole does indeed index some aspect of driving ability.

7.2.2.3 *Item Difficulty*

Item Difficulty is the proportion of participants who answered correctly for a particular question. Thompson and Levitow (1985) suggest an ideal item difficulty lies midway between a chance (25%) and ceiling (100%) score, which for a four-item multiple choice test is 62.5%, or .625. Interestingly, this is almost exactly the mean score across all items (62.3%, or .623), although it varies considerably for individual items, with mean scores ranging from 26.2% to 94.7% (.262 to .947).

7.2.2.4 *Item discrimination*

Biserial correlation coefficients (r_{bis}) were calculated to measure item discrimination, which correlate overall test scores with the response for a particular

test item. In other words, the lowest possible r_{bis}, -1.0, indicates that all those who answered the question correctly obtained low test scores, and the highest possible r_{bis}, +1.0, indicates that all those who answered the question correctly obtained high test scores. This measure was chosen because it is not sensitive to item difficulty, unlike point-biserial or item-total correlations (Attali & Fraenkel, 2000). r_{bis} is reported for all individual items.

The Educational Testing Service, which develops many of the standardized tests for K-12 and higher education in the United States such as the SAT and GRE (Graduate Record Examinations), uses an r_{bis} threshold of +.30 (Zieky, personal communication, cited in (Colbert, 2001), and reviews test items that do not meet this threshold. Some past hazard perception research has also used a point-biserial correlation cutoff of 0.30, for instance (Castro et al., 2014, 2016), and removed any clips not reaching this threshold. All test items had a positive r_{bis} and $r_{bis} > +.30$ for 31 of 40 items, suggesting that the majority of items in the test were adequate discriminators of high and low scorers.

7.2.2.5 *Experience effect*

Mean scores for novice and experienced drivers for each item are reported. Cohen's *d* is reported as a measure of effect size for each item. A single independent t-test compared driver groups across all items, finding that experienced drivers outscored novices overall ($t(182.42) = 3.81$, $p < .001$, $d = 0.56$). This is consistent with the findings from the previous three experiments.

7.2.3 Truncated clip pool

Following the above two analyses (7.2.2.4, 7.2.2.5), the 15 clips with the greatest novice/experience effect size ($d > 0.16$) that also had an r_{bis} above +.30 were

reanalyzed as a single clip pool. An independent t-test confirmed that experienced drivers still significantly outscored novices ($t(174.71) = 5.42$, $p < .001$). For these 15 clips, $d = 0.82$ and Cronbach's $\alpha = .63$.

Table 7-3: Various measures for individual Malaysian clips, collapsing all participants across the three experiments in Chapters 3, 5, and 6.

Item / Clip*	Overall score (%)	r_{bis}†	Nov. driver score (%)	Exp. driver score (%)	Cohen's *d* ‡	Correct responses	Distr. 1	Distr. 2	Distr. 3
MY-M-01*	57.2	.352†	52.6	62.0	0.19‡	107	57	20	3
MY-M-02*	65.8	.342†	54.7	77.2	0.49‡	123	51	5	7
MY-M-03	29.4	.165	31.6	27.2	-0.10	55	97	19	15
MY-M-04	85.6	.575†	84.2	87.0	0.08‡	160	12	7	8
MY-M-05	26.2	.040	25.3	27.2	0.04	49	55	54	28
MY-M-06	64.7	.256	57.9	71.7	0.29‡	121	20	38	8
MY-M-07*	94.7	.547†	90.5	98.9	0.38‡	177	3	2	5
MY-M-08	43.3	.381†	47.4	39.1	-0.17	81	15	64	26
MY-M-09*	87.2	.398†	84.2	90.2	0.18‡	163	16	2	5
MY-M-10	77.0	.437†	74.7	79.3	0.11	144	18	7	18
MY-U-01	34.8	.203	35.8	33.7	-0.04	65	62	41	19
MY-U-02	89.8	.562†	88.4	91.3	0.10	168	11	5	3
MY-U-03*	66.8	.448†	60.0	73.9	0.30‡	125	50	0	12
MY-U-04	95.2	.584†	94.7	95.7	0.04	178	7	0	2
MY-U-05	70.6	.372†	67.4	73.9	0.14‡	132	19	3	33
MY-U-06	60.4	.356†	57.9	63.0	0.11	113	56	18	0
MY-U-07	86.6	.532†	87.4	85.9	-0.04	162	8	13	4
MY-U-08*	72.7	.801†	65.3	80.4	0.35‡	136	21	18	11
MY-U-09*	66.3	.317†	62.1	70.7	0.18‡	124	5	42	16
MY-U-10	35.3	.236	36.8	33.7	-0.07	66	86	22	12

† Indicates $r_{bis} > .300$

‡ Indicates one of the 20 clips with largest novice/experience effect size

* Indicates the 15 clips with the highest novice/experience effect size and $r_{bis} > .300$

Table 7-4: Various measures for individual UK clips, collapsing all participants across the three experiments in Chapters 3, 5, and 6.

Item / Clip*	Overall score (%)	r_{bis} †	Nov. driver score (%)	Exp. driver score (%)	Cohen's *d* ‡	Correct responses	Distr. 1	Distr. 2	Distr. 3
UK-M-01*	76.5	.434†	70.5	82.6	0.29‡	143	15	14	15
UK-M-02	59.9	.294	53.7	66.3	0.26‡	112	27	42	5
UK-M-03*	69.0	.358†	58.9	79.3	0.45‡	129	10	35	12
UK-M-04	54.5	.426†	53.7	55.4	0.04	102	20	34	30
UK-M-05*	82.9	.325†	72.6	93.5	0.58‡	155	11	8	13
UK-M-06	87.7	.322†	88.4	87.0	-0.04	164	17	1	4
UK-M-07*	61.0	.582†	53.7	68.5	0.31‡	114	27	27	18
UK-M-08*	60.4	.356†	52.6	68.5	0.33‡	113	22	30	22
UK-M-09*	44.9	.368†	41.1	48.9	0.16‡	84	47	25	30
UK-M-10	61.5	.440†	58.9	64.1	0.11	115	40	17	15
UK-U-01	70.6	.485†	69.5	71.7	0.05	132	12	22	21
UK-U-02*	46.0	.575†	41.1	51.1	0.20‡	86	66	13	22
UK-U-03	32.6	.249	27.4	38.0	0.23‡	61	41	41	44
UK-U-04	41.2	.431†	38.9	43.5	0.09	77	67	32	10
UK-U-05	55.6	.250	48.4	63.0	0.30‡	104	30	36	16
UK-U-06	72.7	.311†	73.7	71.7	-0.04	136	45	2	4
UK-U-07	51.3	.319†	49.5	53.3	0.08	96	33	24	34
UK-U-08	58.8	.318†	60.0	57.6	-0.05	110	12	7	58
UK-U-09	39.0	.110	40.0	38.0	-0.04	73	47	43	24
UK-U-10*	56.1	.514†	49.5	63.0	0.28‡	105	25	28	28

† Indicates $r_{bis} > .300$

‡ Indicates one of the 20 clips with largest novice/experience effect size

* Indicates the 15 clips with the highest novice/experience effect size and $r_{bis} > .300$

7.3 Discussion

This discussion primarily focuses on the 40-item pool, and the 15-item version of the test will be discussed in 7.3.5.

7.3.1 Experience differentiation

First and foremost, the test as a whole does appear to index some measure of driving ability, given that the previous three experiments all found differences of experience and that experienced drivers clearly outscored novices across all experiments (d = 0.56). This serves as a type of convergent validity; if there were no differences in experience, it would be dubious that the test measured any driving-related skill. As might be expected, individual clips have relatively little power and the efficacy of clips varies greatly within the test. For instance, in nine clips, novices actually outscored experienced drivers, although it should be noted that only one of these clips, MY-M-08, had even a marginal effect size (d > 0.10 in favor of novices), so we may consider the difference negligible for the other clips. Of the remaining 31 clips, the difference in performance was again negligible in 16 (d < 0.20), leaving 15 of 40 clips where experienced drivers outscored novices by any kind of margin.

7.3.2 Item difficulty

We can also see that difficulty for individual items varies considerably, with the hardest item, MY-M-05, having a mean score barely above chance (26.2%), and the easiest, MY-M-07, with a score of 94.7%. Despite being near ceiling, the latter clip was nevertheless a reasonable differentiator of experience (d = 0.38) as its incorrect responses came from almost exclusively novices. While several of the individual clips that differentiate experience would normally be considered too difficult or easy for a standard four-option test (tending towards difficult), it is

interesting to note that the mean score for the 'best' 15 clips mentioned earlier is 63.8%, very close to the overall test score of 62.3% and to the recommended item difficulty of 62.5% for a four-option test.

7.3.3 Reliability

The Cronbach's alpha for the test was .69 which, while reasonable, is lower than similar hazard perception tests (Castro et al., 2014; Scialfa et al., 2011; Vlakveld, 2014; Wetton et al., 2011), although in all cases but one this was after less effective clips had been removed from the item pool. Reliability would need to be considerably higher for a high-stakes test such as one used in licensing (for instance, in the QT-HPT, $\alpha = .93$, although this is with 60 items), but this is perhaps an acceptable result given the clip pool has not been modified in any way. The slight differences between tasks in each experiment may also have influenced reliability; had all 187 participants completed the test under the same conditions, it is possible that reliability would have been higher.

7.3.4 Item discrimination

Thirty-one of 40 clips had an r_{bis} of +.30 or higher, indicating they are good differentiators of high and low performers. Interestingly, in four of these clips novice drivers outscored experienced drivers, suggesting that factors other than experience contribute to "What Happens Next?" performance. While experienced drivers likely make up the majority of high performers and vice versa for novices, there is inevitably some overlap between the groups and individual differences certainly play a role. For instance, UK drivers outscored Malaysian drivers by a considerable margin in both cross-cultural experiments, which certainly contributes to the variability in performance among participants; in fact, the difference between UK and

Malaysian drivers was greater than the difference between experienced drivers and novices, indicated by the main effect sizes in the accuracy analyses (see 8.3.1 and sub-sections for more detail and discussion). Furthermore, the item discrimination measure adds a level of analysis that could not be achieved when investigating driver experience alone; as stated above, high-performing novices and low-performing experienced drivers certainly exist, and the biserial correlation coefficient allows items to be eliminated that do not discriminate performance, rather than purely experience.

7.3.5 Summary

Overall, the results of these experiments are certainly promising and suggest that this version of the "What Happens Next?" test provides a reasonably valid and reliable test of driving ability, even in its initial form and before refining the clip pool. Refining the clip pool also results in improvements in experience differentiation. After selecting the 15 clips with the largest novice/experience difference that were also adequate discriminators of high and low scorers, the effect size was slightly higher ($d = 0.82$, up from $d = 0.56$); although this is still considerably lower than in the QT-HPT where $d > 1.18$ for all 15-item tests. Internal reliability measured by Cronbach's α decreased slightly to .63 from .69, although this is perhaps unsurprising given the smaller clip pool and the nature of a binary response format. While the truncated clip pool still does not meet the standard of a high-stakes test such as one to be used in licensing, the results are certainly promising should one specifically aim to create a high-stakes test using the "What Happens Next?" multiple choice paradigm.

CHAPTER 8

GENERAL DISCUSSION

8.1 Aims of this thesis

The research presented in this thesis had two main aims. The first was to examine hazard perception in a cross-cultural context, namely how a driver's environment might affect their hazard perception ability. The experiments described have therefore used videos filmed in both the UK and Malaysia (with the exception of Chapter 4, which used only Malaysian videos), and where possible, participants from both the UK and Malaysia (Chapters 2, 3, and 5), to investigate any cultural interactions that arose. The general conclusions from this strand of research are discussed in Section 8.3. The second goal of the work in this thesis was to validate a hazard perception test for potential use in driver licensing, with a particular view towards implementation in lower-income, developing countries. While the reaction time test has been researched in many countries previously (Australia, Canada, the UK, Israel, and Norway, to name a few) , these have all been high-income countries with relatively low accident rates, and this thesis describes the first published research conducted on hazard perception in both a developing country and a cross-cultural context (Lim et al., 2013, 2014). The general conclusions and practical considerations for hazard perception testing in developing countries are discussed in Section 8.4.

8.2 Summary of data presented

This thesis employed two video-based hazard perception tasks: the traditional reaction time task that requires drivers to respond as soon as they detect developing hazards (Chapter 2), and the "What Happens Next?" test, a predictive task that requires drivers to choose or describe hazards that would have happened next (Chapters 3 – 6). We concluded that the reaction time task may have validity issues in countries with high accident rates, where drivers appear to be desensitized to hazards (Chapter 2); ironically, the very countries that arguably need a hazard perception test the most. The "What Happens Next?" test was examined as a possible alternative and found to differentiate experience in drivers from both the UK and Malaysia (Chapter 3); therefore all subsequent thesis chapters (4 – 6) use this paradigm. All subsequent chapters also use the same videos, with the exception of Chapter 4. Chapter 4 was also the only chapter to additionally incorporate a free response format, which was compared directly to the multiple choice format (the sole format used in all other "What Happens Next?" experiments in this thesis). Results indicated that the free response format was more powerful than multiple choice, although free response was noted to be less feasible from a practical perspective, since its responses require interpretation and a potential licensing test would require large-scale scoring. However, while the multiple choice format does appear statistically weaker, it is certainly possible to create an acceptable test using this format: the UK set of videos used in Chapter 3 differentiated experience, and the Malaysian and UK sets of videos differentiated experience equally well in subsequent chapters (Chapters 5, 6). Furthermore, the results of these three chapters (3, 5, and 6) may indicate that recruiting drivers with greater levels of experience improved the test's experience differentiation. Chapter 6 also indicated that reducing video quality did not affect test

effectiveness, a potential concern for online hazard perception testing. Finally, an item analysis (Chapter 7) suggested that the "What Happens Next?" videos used throughout most of this thesis (i.e. Chapters 3, 5, 6) had acceptable reliability and validity when taken as a whole test, although the capacity for individual videos to differentiate experience seemed inconsistent from experiment to experiment.

On the more theoretical side, we concluded that culture certainly plays a role in how drivers view hazards; for instance, drivers accustomed to more dangerous environments (i.e. Malaysians, in the context of this thesis) not only have increased thresholds for identifying hazards (Chapter 2), they also predict near-imminent hazards less accurately (Chapters 3, 5). Chapter 4 tentatively indicates that Malaysian drivers have and/or develop different levels of hazard awareness to UK drivers, although this conclusion is based on an indirect comparison with a previous "What Happens Next?" UK study (Jackson et al., 2009). Overall, however, hazard perception ability of an individual seems relatively stable across cultures, as UK and Malaysian drivers reacted fairly similarly to hazards filmed in both their home and non-home country (Chapter 2), and other than some minor effects in Chapter 2 (see 2.4.1) there were no cultural interactions; for instance, the Malaysian set of hazards were universally easier to predict for all drivers regardless of origin (Chapters 3, 5).

8.3 Influences on hazard perception

8.3.1 Hazard perception in a cross-cultural context

In Chapter 3, we found that Malaysian drivers were slower to identify hazards than UK drivers. Other analyses suggested this was due to a higher threshold for identifying hazards; for instance, in the same experiment, Malaysian drivers identified

less hazards overall than UK drivers, for both pre-defined and non-pre-defined events. They also fixated hazards at the same time as UK participants, raising the possibility that non-responses were not due to a failure to visually detect the event, but rather a failure to perceive it as hazardous. We concluded that the differing hazard thresholds likely stem from driving in different environments; the higher accident rate and more hazardous environment in Malaysia likely lead to drivers in this environment being desensitized to hazards.

Altogether, the results of Chapter 2 indicated significant cross-cultural differences in how dangerous Malaysian and UK drivers perceive hazards to be, providing the main impetus to explore other test paradigms. The results of Chapters 3, 4 (to a certain extent), 5, and 6 indicate that the "What Happens Next?" test is a viable alternative for hazard perception testing, particularly in Malaysia or other countries where drivers may be similarly desensitized to hazards. Additionally, it is arguably a more powerful test than the traditional reaction time paradigm, since the same videos failed to differentiate experience in the traditional paradigm (Chapter 2) but did in the "What Happens Next?" test (Chapters 3, 5, 6), despite being selected specifically for the reaction time test.

However, although hazard criterion does not play a role in the "What Happens Next?" test, UK drivers outscored Malaysian drivers in all relevant experiments (Chapters 3 and 5). Furthermore, they outscored them by a wider margin than experienced drivers outscored novices, indicated by the main effect sizes in the accuracy analyses (Chapter 3: $\eta^2_\rho = .094$ for country, $\eta^2_\rho = .057$ for experience; Chapter 5: $\eta^2_\rho = .161$ country, $\eta^2_\rho = .093$ experience). This suggests that although experience certainly plays a role in hazard perception ability, other factors also contribute significantly. While a driver's country of origin is an obvious factor, it is

difficult to pinpoint a precise cause beyond general 'environment'; for instance, while the higher rate of accidents and different driver training in Malaysia has been discussed (also see Section 8.3.2), its infrastructure and enforcement almost certainly contribute as well, if to lesser degrees.

8.3.2 Driver training

One obvious difference between Malaysian and UK drivers is the driver training received in their respective countries; almost all UK drivers tested would have practiced for, and passed, a hazard perception test in order to obtain their license. Novices in particular would have undergone their training less than a year prior to participating in the experiment, possibly resulting in artificially higher scores on the reaction time test (Chapter 2) in particular, as they were practicing for the very same test. These immediate practice effects may partly explain the lack of experience differentiation among UK drivers in Chapter 2's reaction time test, given that similar tests have differentiated experience among UK drivers in the past (Mckenna & Crick, 1991; Watts & Quimby, 1979), although this is still debatable since other studies conducted before the UK hazard perception test's implementation have failed to find experience differences (Chapman & Underwood, 1998; Crundall et al., 2003). However, while UK drivers' faster response latencies in Chapter 2 may be partly attributed to having more specific and recent training than Malaysian drivers, training effects have been noted to generalize (Fisher et al., 2006; Isler, Starkey, & Sheppard, 2011; McKenna et al., 2006), and UK participants were also better able to predict hazards in both the UK and Malaysia in both Chapters 3 and 5, suggesting that specific training effects are unlikely to be a major contributor to the performance difference between UK and Malaysian drivers.

It seems possible that in the UK, the major benefit of the hazard perception test requirement stems not from the test itself, but from the emphasis it places on different aspects of driver training. For instance, during lessons, UK driving instructors regularly caution their students to keep watch for potential hazards and point out any relevant precursors, a technique akin to commentary training, which has been shown to improve hazard perception among other driving behaviors (Castro et al., 2016; Isler et al., 2009; McKenna et al., 2006; Wetton et al., 2013). This explicitly teaches them common hazards and potential precursors, enriching any related schema. While this is done with the immediate goal of preparing their students for the hazard perception test, it also instills a hazard-aware mindset in new drivers; effectively, associating driving with keeping alert for potential hazards. In contrast, in Malaysia the focus remains largely on the mechanical aspects of driving, for instance changing gears and parallel parking, and less if any emphasis is placed on watching for other road users. While some common driving situations are incorporated into the written exam, this does not interact with the on-road test at any point; this likely results in drivers acquiring declarative rather than procedural knowledge, which, as the results of Chapter 5 seem to suggest, may be of limited use.

Another interesting finding is that UK drivers' behavioral data in Chapter 2 tended to be more consistent than those of Malaysian drivers. For instance, there is much less variability among UK drivers in reaction time, accuracy, and hazard ratings. This was not the case in Chapter 3, where out of the three primary measures of accuracy, confidence, and hazard ratings, only hazard ratings showed less variability among UK drivers (consistent with Chapter 2), while variability in accuracy was somewhat equal and there was actually more variability among UK drivers in the confidence measure. More interestingly, the consistency in UK drivers'

responses is mostly evident in the behavioral data; in contrast, the eye tracking data shows that variability was fairly comparable between both groups of drivers for all three of the major measures (time to first fixation, mean fixation duration, and horizontal search).

The greater consistency among UK drivers in accuracy and reaction time may be another effect of training; because all UK participants had practiced for and taken the hazard perception test, this effectively acts as hazard calibration, giving test-takers a more uniform idea of what constitutes a hazard (accuracy) and an appropriate time to respond (reaction time). In Malaysia on the other hand, there is no such calibration. While it is highly unlikely that *no* driving instructors in Malaysia caution their students to watch for hazards, this is entirely up to the instructor and therefore inconsistent; furthermore, individual differences for hazard tolerance also exist in instructors and likely to a greater extent than in the UK, creating even more variation in a given driver's learning experience in Malaysia. Some theory test questions in Malaysia do involve various driving situations and common responses, but this is naturally no substitute for on-road or video depictions of hazards. It is difficult if not impossible to achieve a uniform definition of hazardousness with language alone, as descriptions of common road situations may be interpreted differently by different individuals; one person's idea of another road user "driving too fast" or "getting too close" could be perfectly acceptable to another driver with a higher tolerance for danger and/or better judgement of distance.

The eye tracking data on the other hand showed similar consistency among both UK and Malaysian drivers; neither group exhibited more or less variability than the other. This observation mirrors the results reported in Chapter 2, where there were fewer differences between UK and Malaysian drivers in the eye tracking data

compared to the behavioral data. As discussed in several chapters (2.4.3, 3.4.3), it is possible that behavioral differences manifest faster than their corresponding visual strategies, although past research on this point seems somewhat mixed, with some studies reporting eye tracking differences without corresponding behavioral differences (Chapman & Underwood, 1998; Crundall et al., 2003) and some reporting relatively few scanning differences but finding behavioral differences (Borowsky et al., 2010).

8.3.3 Driving environment

Both empirical (Chapter 2, Section 2.3; see reaction time, accuracy, and eye tracking measures) and anecdotal data indicate that drivers in Malaysia are desensitized to hazards; for instance, during all three experiments involving both countries (Chapters 2, 3, and 5), many Malaysian participants made statements along the lines of "that's normal for here" regarding the Malaysian videos, while UK participants often commented on the higher hazardousness of the events in Malaysia. As discussed earlier (see 2.4.4, 3.4.2, 8.3.1), this desensitization almost certainly contributes to Malaysian drivers being generally slower to identify hazards, as events need to reach a higher level of danger before Malaysian drivers would identify them as hazardous (Chapter 2). This corresponds with the findings of Wallis & Horswill (2007), who reported that trained and experienced drivers responded to hazards both faster and more often than untrained and novice drivers respectively, suggesting a response bias largely drove the differences in hazard perception skill. Notably however, there are considerable individual differences among drivers' thresholds for hazardousness, as there were a number of Malaysian drivers who responded with similar speed to UK drivers (also leading to greater variance among them, as discussed in 8.3.2 above).

It is interesting however that Malaysian drivers also predict hazards with lower accuracy compared to UK drivers (Chapters 3, 5). As discussed in Chapter 3 (see 3.4.2), perhaps a hazardous driving environment negatively impacts one's hazard perception ability, contributing at least in part to higher accident rates; this is especially concerning because it could potentially create a self-perpetuating cycle, where a hazardous environment causes people to develop riskier driving habits that in turn cause more accidents or near-accidents, and so on. It is possible that because of the higher hazard frequency, drivers in Malaysia deploy their attention more evenly but also more shallowly, which while more suited to keeping track of simultaneous potential hazards, would make drivers less able to predict any one particular hazard (see 3.4.4). The results of Chapter 4's free response experiment suggest this somewhat; Malaysian drivers identified imminent hazards, their locations, and what might happen next with equal accuracy, in contrast to Jackson et al.'s 2009 study where UK drivers' accuracy dropped with each question (hazard, location, and prediction respectively). However, we also would have expected to find Malaysian drivers' visual search patterns to reflect this strategy in Chapters 2 and/or 3, although again, the general lack of differences between even novice and experienced drivers may mean that visual strategies had not yet developed fully; alternately, perhaps other eye tracking metrics such as those measuring scan paths might have revealed differences. If a strategy of shallow attentional deployment does indeed exist, it seems somewhat detrimental to performance in both the reaction time test and the "What Happens Next?" test, but may be a necessary adaptation in Malaysia. In fact, in the Chapter 4 results described above (also see 4.2.2.1), novice drivers had the highest scores on the hazard question but the location and prediction questions were equal; in contrast, experienced drivers had similar scores on all three questions,

suggesting that a strategy of more diffuse attentional deployment might develop with greater exposure to the Malaysian driving environment.

The possibility that Malaysian drivers deploy their attention more widely but consequently more shallowly is an interesting one. In Chapter 4 (see 4.1) we discussed the possibility that novices found it harder to generate potential hazards on their own compared to experienced drivers in the "What Happens Next?" free response format, but were relatively better at choosing a hazard if potential precursors had already been highlighted for them, as in the multiple choice format. Crundall et al. (2012) reported that novices were more likely to miss fixating behavioral prediction precursors (BP; where the hazard and precursor were the same), which suggests they may be less aware of the potential danger posed by these precursors. While this may explain why the free response format differentiates experience better than the multiple choice format, the free response format is also more statistically powerful than multiple choice, so at present it is difficult to separate these effects.

The difference between generating one's own responses and choosing from possible precursors/hazards somewhat parallels a general difference between the UK and Malaysian videos. One might argue the nature of the multiple choice task changes subtly between UK and Malaysian clips. Because the average UK clip has less potential hazards and precursors than a Malaysian clip, the task in a UK clip arguably amounts to detecting what is often the sole precursor. In contrast, in a Malaysian clip, predicting a hazard might be more similar to choosing which of several highly salient precursors develops into an eventual hazard. While this does not appear to affect experience differentiation when the experienced driver group has held their licenses for a longer time (Chapters 5 and 6), it may make a difference with a slightly less experienced group (Chapter 3), where only the UK videos differentiated

experience. Somewhat resembling the difference between UK and Malaysian (and free response and multiple choice) clips, (Vlakveld, 2014) reported two tasks where participants first had to report the highest priority potential hazard after watching a clip, then reported all potential hazards while watching a clip and later chose the one with the highest priority. The second task showed less experience differentiation, also suggesting that novice drivers were relatively better at picking the highest priority precursor out of a pool of several.

8.3.4 Skill transferability

Research conducted by Wetton et al. (2010) has suggested that hazard perception skill is highly transferable, finding high correlations between Queensland drivers' response latencies to hazards filmed in two different parts of Australia and the UK. Chapter 2 confirmed this transferability with more disparate driving environments and used participants from both countries, allowing for potential cultural interactions. The most notable cultural interaction was that drivers detected more pre-defined hazards from their home country, suggesting that environmental familiarity primes them to react appropriately once a hazard appears; this is likely due to having richer mental models in familiar environments (Underwood et al., 2002). Apart from this, drivers showed the same general response pattern for most measures; for instance, although Malaysian drivers detected less hazards than UK drivers as a whole, all drivers responded to UK unmatched hazards the least of any clip group. We concluded that overall, while certain nuances of hazard perception are affected by context, drivers view and respond to hazards similarly regardless of their home driving environment. In other words, relative danger stays the same regardless of context.

Chapters 3 and 5, the two cross-cultural "What Happens Next?" chapters, provided further evidence for hazard perception skill being highly transferable, as none of the analyses indicated any cultural interactions. As in Chapter 2, drivers responded similarly regardless of their country of origin; for instance, all drivers found Malaysian hazards easier to predict than UK hazards. Along with predictive accuracy, experience differentiation across cultures also remained consistent; in Chapter 3 only the UK set of videos differentiated experience, while in Chapter 5 both sets of videos differentiated experience equally well, and this was the case for both Malaysian and UK drivers. These observations are of course based on quantitative data restricted to the laboratory, and qualitative research, on-road observation or simulator work may uncover many cultural differences that are beyond the scope of these experiments; however, hazard perception skill at least remains fairly consistent.

It is perhaps unsurprising that drivers view and respond to hazards similarly across different cultures; while the same action can often be perceived differently in different cultures, hazards are arguably an exception because a given hazard generally presents a certain amount of danger with little variation. For instance, higher speeds are almost always considered more hazardous, as are larger vehicles; it is difficult to think of a situation where a leisurely bicyclist might be more dangerous than a speeding truck. Furthermore, driver licenses are somewhat transferable; for instance, holding a license from one country often allows one to drive in many other countries, and one country's license can often be converted to another's without having to take a new licensing exam. These policies assume at least some degree of standardization among licensed drivers.

However, while hazards may be perceived similarly on a relative scale, culture and/or driving environment appear to affect hazard perception skill considerably on

an absolute scale, which has certain practical implications; namely, that drivers from a country with high accident rates may exhibit lesser hazard perception skill compared to drivers from countries with better road safety records. Perhaps reflecting this, only certain countries can convert driver licenses to other countries'. One obvious example is that licenses obtained in Malaysia cannot be converted to UK licenses; while drivers can convert say, an Australian license, drivers with only a Malaysian license must take the UK licensing exam to obtain a UK license. Based on the data presented in this thesis, these policies may be well-founded.

One interesting possibility this raises is how much a driver's current road environment affects their hazard perception skill. In the experiments presented in this thesis, all participants had learned to drive in either the UK or Malaysia, and with few exceptions, had also lived in that country for their entire life. We would therefore expect the observed levels of hazard perception skill to be a reasonable representation of the general population, particularly in the UK, due to greater standardization of driver training and the hazard perception test providing a more uniform idea of what constitutes a hazard (discussed in 8.3.2 above). The data presented in this thesis has established that UK drivers trained and living in the UK generally predict and react to hazards better and faster than Malaysian drivers trained and living in Malaysia; however, what is not clear is whether UK drivers would maintain the same level of skill if they relocated to Malaysia, or vice versa for Malaysian drivers relocating to the UK. While various types of training have been shown to improve hazard perception (Castro et al., 2016; Isler et al., 2009; McKenna et al., 2006; Wetton et al., 2013), none have been conducted cross-culturally. This would be especially interesting as it could shed some light on how stable hazard perception abilities are,

and give some insight into the separate effects of training and the surrounding environment.

8.4 Validating a hazard perception test for driver licensing

The second primary goal of this thesis was to explore a test for potential use in driver licensing, as a way of improving road safety in developing countries with high accident rates such as Malaysia. These tests, and other driving-related tasks, are generally validated by a correlation between driving experience and task performance, with experience considered a proxy for driver safety. While a direct link between a driver's road safety record and test performance would be ideal, this is generally not reliable without very large sample sizes (Horswill & McKenna, 2004).

Similar to many other countries, novice drivers in Malaysia have disproportionately poor road safety records, making driver experience a valid proxy for safety; therefore in this thesis, we have also used experience as a proxy. Ideally, any test paradigm that may be eventually used in the licensing exam (and even more ideally, the test itself), will also be validated by directly correlating driver safety and test performance in Malaysia, although this is work beyond the scope of this thesis.

8.4.1 Paradigm comparisons

The reaction time test is the de facto test for hazard perception ability and is used in the driver licensing exam in the UK and Australia. Many researchers have found response latency on this test to correlate with driving experience and/or accident involvement (Horswill et al., 2008; Scialfa et al., 2011; Wallis & Horswill, 2007; Wetton et al., 2010), although other studies have failed to find this experiential difference (Chapman & Underwood, 1998; Crundall et al., 2002; Sagberg &

Bjørnskau, 2006). More directly, several prospective studies have indicated a link between response latency and driver safety; for instance, Boufous et al. (2011) reported that drivers who had failed the New South Wales hazard perception test multiple times had higher crash rates, Wells et al. (2008) found that UK drivers who had taken a hazard perception test had slightly reduced accident rates for some categories of accidents, and most recently, Horswill, Hill, et al. (2015) reported that drivers' response latencies in the QT-HPT corresponded with later accident rates.

However, these studies have all been conducted in Australia or the UK, and at the time of this work, the reaction time paradigm had not yet been studied in Malaysia; Chapter 2 describes the first use of the reaction time paradigm in a developing country. As discussed in Chapters 2 and 3, despite its advantages in the UK and Australia, the traditional reaction time test faces some validation issues in Malaysia and likely, other countries with similar accident rates. For this reason, this thesis has focused on the "What Happens Next?" test; however, this paradigm also faces its own set of advantages and disadvantages. This section therefore compares the two paradigms.

8.4.1.1 *General considerations for hazard perception testing*

Unlike the reaction time paradigm (Chapter 2), the "What Happens Next?" task differentiated between experienced and novice drivers in all three studies (Chapters 3, 5, and 6) using the same clips. This is especially encouraging given that two of these studies, Chapters 3 and 5, used both UK and Malaysian participants, suggesting the task differentiates experience while remaining largely culture-agnostic. Furthermore, while both paradigms utilized the same set of videos, only the predictive task differentiated experience, which is notable given that the videos were selected

specifically for the reaction time task, raising the possibility that the predictive paradigm may be a more powerful differentiator of experience.

While some of these points have already been discussed in earlier chapters (2.4.6, 3.4.1), there are some issues with the reaction time paradigm that a predictive paradigm circumvents. For instance, when using a button press paradigm one cannot be sure that participants are responding to the same hazard defined by the researchers or a different hazard altogether. It is also easier to deceive a button press paradigm than a predictive task; for instance, while the UK test attempts to circumvent cheating by failing participants who respond at particularly high rates, some participants may attempt to press the response button at a speed just below the one which would be flagged. Both these issues have been compensated for in various ways, such as using a touchscreen, asking drivers to click on the hazard with a mouse, or asking drivers to verbally identify the hazard (Chapter 2; (Lim et al., 2013; Scialfa et al., 2011; Wetton et al., 2010, 2011), although the former solutions may be more suitable for mass testing since the last requires responses to be interpreted. Finally, some exceptional participants may also detect a hazard and respond before its window opens, effectively penalizing them for responding early.

One might argue that scoring the reaction time test is more subjective because hazard onsets have to be defined; when does a precursor become a hazard? An offset must also be defined if using a hazard window, which adds another layer of subjectivity. On the other hand, the "What Happens Next?" task requires more pre-test preparation because the four options for each video must be decided and worded appropriately. Furthermore, since the test requires some level of literacy, it may be disproportionately challenging for those with lower English proficiency, which is an especial concern in countries like Malaysia. While translating the text is certainly an

option and may in fact be necessary in Malaysia, since the written theory test is offered in both English and Bahasa Malaysia, it would considerably increase preparation time, especially during the validation phase as videos would need to be validated in both languages.

8.4.1.2 *Hazard criterion and desensitization*

Initially we hypothesized that Malaysian drivers would either be desensitized or highly responsive to hazards compared to UK drivers, and Chapter 2 presents fairly clear evidence for the former conclusion. While it is possible that drivers in other developing countries may be highly sensitive to hazards due to greater exposure, it seems more likely that they will display similar desensitization to Malaysian drivers given the similar crash records, although of course this will first need to be confirmed. A reaction time test therefore presents a validation issue in Malaysia, as individual differences in hazard criterion (*hazard classification*, the third component of hazard perception; (Wetton et al., 2010) are much larger than in Australia or the UK and thus greatly influence response latency. However, Crundall and Chapman (2014) reported that modifying test instructions for UK drivers reduced non-hazard responses without reducing *a priori* hazard responses, so it certainly seems possible that a reaction time paradigm could be successfully calibrated. This provides an evident avenue for future work.

One might argue that the issue of hazard criterion does not make the reaction time test unworkable and in fact, may actually be a point in its favor over the "What Happens Next?" task. After all, desensitization to hazards almost certainly contributes to higher accident rates, and part of the reaction time test's utility in the UK has been raising new drivers' awareness of hazards; in other words, perhaps the issue we would avoid by using a different licensing test is precisely the one that needs

to be addressed. As discussed earlier (8.3.2), the implementation of the hazard perception test in the UK has certainly had some advantages, the foremost among them being heightened awareness of hazards among new drivers. It is reasonable to imagine that implementing a test that uses the reaction time paradigm would eventually have similar trickle-down effects in Malaysia. However, as we will discuss later in more detail, road safety is a complex issue with many different factors to consider. While this thesis mainly addresses the cognitive aspects of driving, other factors undoubtedly play a role such as enforcement, road infrastructure, and attitudes toward road safety, all of which interact. Introducing a new component to the licensing exam represents a major change that comes with its own set of challenges, and while one of the ultimate outcomes should indeed be greater sensitivity to hazards, this is arguably a side effect of the primary goal, which is improved road safety. What is certain is that introducing a new test *and* trying to change drivers' attitudes and ideas about hazards will be challenging at best, and it may be especially difficult if the success of the first partly relies on the second. Therefore, it might be wise to focus on changes that are less sweeping (and theoretically offer less resistance), but have the potential to have the same effect; introducing a test that does not rely on changing people's ideas of hazards may be a smaller and therefore more viable step.

8.4.1.3 *Empirical support*

One potential concern with adopting the "What Happens Next?" test is that it is a less established paradigm compared to the reaction time test, which may pose some potential difficulties in attaining government buy-in. While the reaction time test has been researched in Australia, the UK and even Singapore (Yeung & Wong, 2015), as of late 2015, there has been little published data on the "What Happens

Next?" test (Jackson et al 2009, Castro et al 2014, Lim et al 2014), with (Castro et al., 2016; Crundall, 2016) forthcoming, and of those five, only one study uses the multiple choice paradigm. Arguably, neither paradigm has been researched extensively in developing countries, but this is certainly a practical consideration.

8.4.2 "What Happens Next?": Methodological considerations

This section discusses some methodological considerations of adopting the "What Happens Next?" test. Note that we focus exclusively on the multiple choice test rather than free response, due to its practical viability for mass testing. When comparisons are made between experiments, unless otherwise stated, they concern the three chapters that used the same set of videos (Chapters 3, 5, and 6).

8.4.2.1 *Statistical power*

In Chapter 4, we discussed how the free response format of "What Happens Next?" was statistically more powerful than the multiple choice format, but the need for response interpretation makes it undesirable as a mass testing paradigm. Similarly, the reaction time test offers more statistical power than "What Happens Next?", since, like the free response format, performance is measured on a continuous scale while the multiple choice format uses a binary scale. As an example, the current UK licensing exam uses 15 hazards in 14 videos, and test-takers score 0 – 5 points for each hazard depending on how early in the hazard window they respond, allowing a maximum score of 75 points (UK Department for Transport, 2016),. The current Queensland licensing test also uses 15 hazards/videos. A multiple choice paradigm, on the other hand, allows only a 0 or 1 score on each question and only one scoring opportunity per video, giving a maximum of 15 points with 15 videos, a much lower score range. It is therefore possible that more than 15 videos will be necessary to achieve reasonable levels of reliability if the "What Happens Next?" test is used,

adding to both the time spent by drivers to sit the test and the amount of time taken to prepare it.

8.4.2.2 *General consistency*

Several correlations were conducted across the three experiments that used the same "What Happens Next?" videos (Chapters 3, 5, and 6; see 7.2.1). A Pearson's correlation was run to compare accuracy (% of participants who answered a video correctly), experience differentiation (measured by effect size in Cohen's *d*), and balance of distractor options (measured by the goodness of fit chi-square test value comparing the three distractor options). Accuracy was highly correlated across all three experiments (all $rs > .74$, all $ps < .001$), suggesting that the relative difficulty of each video remained constant regardless of the conditions under which participants viewed it. Chi-square test value also significantly correlated although not as strongly as accuracy (all $rs > .45$, all $ps \leq .003$), suggesting that the distribution of responses among distractor options also stayed reasonably constant.

It should be noted that because the chi-square value measures only how unevenly responses were distributed among the three distractors, this does not necessarily mean that responses were skewed towards the *same* distractors every time; only that they were distributed among all three similarly. For instance, in one experiment, responses could have been distributed like so: Option 1, 30 responses; Option 2, 20 responses; Option 3, 60 responses. In another experiment, they could have been distributed the same way but among different options, for example: Option 1, 20 responses; Option 2, 60 responses; Option 3, 30 responses. In both cases the chi-square value would be the same. However, in practice, this seems less likely than the same distractors drawing similar responses, and a quick glance at the response distribution confirms this is generally not the case.

We can therefore conclude that accuracy and distractor responses stay roughly similar across experiments; in other words, overall response distribution is somewhat consistent. However, this is not the case for experience differentiation. There was no significant correlation in a video's effect size across any of the three experiments, although Chapters 5 and 6 showed a non-significant negative correlation, unlike the other correlations which were positive.

8.4.2.3 *Distractor options*

In a classroom multiple choice test, it is fairly intuitive that more plausible distractor options lead to a higher quality question; if students can easily eliminate two options in a four-option test, this gives them a 50% chance of guessing the correct answer, effectively doubling the usual 25%. In the "What Happens Next?" test, it stands to reason that if one distractor option was chosen disproportionately more or less often than the others, this might signify a poorer quality item. It is debatable whether this was the case across these experiments, as none of the three experiments showed a significant correlation between how well a clip differentiated experience (measured by effect size, Cohen's *d*) and how balanced its distractor options were (measured by chi-square test value). However, it is perhaps notable that in every experiment, there was a negative correlation between these two variables (r = -.258, -.096, -.188 for Chapters 3, 5, and 6 respectively); in other words, clips that differentiated experience better tended to have distractor options that were more evenly distributed. This might tentatively suggest that if there is indeed a link, plausible distractor options contribute to a clip's experience differentiation, but the correlation is fairly weak and likely not a high priority factor.

8.4.2.4 *Online testing*

The results of Chapter 6 indicate that the "What Happens Next?" test remains effective even after reducing the quality of the videos used in the test, as videos that had been considerably degraded still differentiated experience effectively. Indeed, it seems possible that degradation may actually improve test effectiveness, given that experience differentiation was greater in Chapter 6 (where medium and low quality videos were used) compared to Chapters 3 and 5 (where only high quality videos were used); see Section 6.4.1.2 for details. However, this particular conclusion is tentative and will certainly require further testing.

More importantly, the results of Chapter 6 have implications for online hazard perception testing. Because some drivers may have limited internet speed and/or access, it may sometimes be necessary to use small file sizes and therefore lower quality videos in online tests. This was indeed the case in the Queensland Transport Hazard Perception Test, (QT-HPT: Horswill, Hill, et al., 2015). Horswill, Hill, et al. reported that test scores on the QT-HPT were linked to both retrospective crashes and prospective crashes within a year of test-taking, confirming that the QT-HPT was still an effective index of crash risk despite the reduced quality of the videos used in the test. Given that both Horswill et al. and Chapter 6 report experience differentiation with reduced quality videos, this is promising news for online hazard perception testing. The fact both studies used different hazard perception test paradigms but reported similar results is especially encouraging, as this suggests that reduced quality videos remain effective across the board.

8.4.2.5 *Clip selection and creation*

It is difficult to ascertain qualities that make a clip suitable for differentiating experience, especially because of the inconsistencies mentioned earlier in Section

8.4.2.2. It may be that individual clips have less power and are perhaps less consistent in how well they differentiate experience, but their cumulative effect results in an entire test that does reliably differentiate experience. The analyses described in Chapter 7 certainly suggest this, as a truncated 15-item pool showed better experience differentiation than the 40-clip pool, although internal reliability decreased (see 7.2.2.5 and 7.2.3 for more detail). Perhaps, then, a better strategy for test creation is to focus on excluding poor differentiators, rather than increase the number of good differentiators. As discussed in 8.4.2.1, it is also possible that the experiments described in Chapters 3, 5, and 6 lack the statistical power to effectively distinguish individual clips, given that the test uses a binary measure of performance. While the nature of individual clips certainly plays a role (Sagberg & Bjørnskau, 2006), it is not feasible with the present data to reliably distinguish characteristics of effective "What Happens Next?" clips, beyond the general strategy of choosing clips with appropriate precursors (Wetton et al., 2011).

8.5 Other considerations for driver safety in Malaysia

Road safety is a complex issue with many different factors to consider. While we have discussed some of the cognitive aspects of driving, specifically those related to hazard perception, other factors also play a major role. For instance, anger, fatigue, and seat belt use have been studied among Malaysian drivers (Hauswald, 1997; Mohamed et al., 2012; Sullman et al., 2014, 2015). A glance at the recent research reports of the Malaysian Institute of Road Safety Research (MIROS) may give us a general idea of the government's road safety priorities; there are a large number of reports concerning motorcyclists in particular, but other topics include crash injuries, red light violations, automated enforcement systems, and traffic offenses during

country-wide holiday periods such as Chinese New Year and Hari Raya. The emphasis on motorcyclists likely stems from the fact that motorcyclists generally cause a disproportionate number of accidents; additionally, motorcycles make up a significant amount of traffic in Malaysia (47%; Toroyan, 2013), likely adding to the already high accident rate.

The focus on automated enforcement systems also highlights that this is a significant problems in Malaysia; consistent enforcement, or rather lack of, is a recurring issue in the country in many areas beyond road safety. As a result, traffic violations occur on a regular basis. In fact, the vast majority of the Malaysian videos filmed for this thesis include several if not dozens of minor breaches. Some of the violations that occur include illegally parked vehicles, cars stopped in the middle of the road, overtaking in the wrong lane, and jaywalking pedestrians, to name a few. While some automated systems have been implemented, their effectiveness is limited to certain areas; for instance, they are used primarily for traffic light and speeding violations, but cannot be easily applied to any of the common aforementioned violations. Furthermore, even with automated systems, enforcement is inconsistent and delayed; a driver who regularly exceeds the speed limit on the way to work every day will possibly receive one speeding ticket in a month, notification of which often arrives several months after the violation.

Another issue that can be seen in some experimental videos is one of infrastructure. Although still considered a developing country, Malaysia has undergone rapid industrialization, particularly in and around its capital, Kuala Lumpur. The expansion of Greater Kuala Lumpur has resulted in city planning being somewhat reactive at times, making it difficult for drivers to navigate the city. For instance, signage is often confusing or lacking altogether. This also applies to road

markings; lane markings for instance are often inconsistent, as can be seen in some of the videos, and one common issue is a lane that has been marked as straight-ahead later becomes turn-only. Another example is generally unforgiving road planning, where if a driver takes a wrong turn, they might drive twenty minutes before finding a route that lets them turn around. These idiosyncrasies all contribute to create an environment where some road users accept a certain amount of danger as second to convenience; for instance, while not common per se, it is certainly not unheard of to see drivers or particularly motorcyclists driving the wrong way on a road in order to avoid a long detour.

Another concern that is implicit in most of the issues described in this section, as well as parts of this thesis, is the population's attitude toward road safety. As discussed earlier (8.3.1, 8.3.3), it is possible that a hazardous driving environment leads to driving behaviors that encourage greater danger, which leads to a more hazardous environment, and so on. This attitude was indeed observed in many Malaysian participants, where drivers commented that the kind of hazards presented in the videos were "normal for here." Again anecdotally, many Nottingham staff who grew up and obtained their drivers license in the UK have also commented that their driving habits have deteriorated significantly since moving to Malaysia. (Incidentally, this has been indirectly confirmed by the reactions of drivers in the UK when they have returned home and driven there!) Indeed, most of the factors described above likely contribute to a generally lackadaisical attitude toward driving safety. When roads are so poorly maintained cars often get nicks and scratches, and drivers know they will likely not receive any penalties for minor violations, they are likely to adopt a more careless attitude; "what's one more scratch, after all?"

The issues discussed above are a few of many that add to the complexity of improving road safety in Malaysia. Given the government's goal of reducing road fatalities by 50% by 2020, it seems clear that major and rapid action is necessary to achieve these targets. The data presented in this thesis lays the groundwork for a licensing test that we hope will be of use in Malaysia and similar countries aiming to improve road safety.

REFERENCES

Attali, Y., & Fraenkel, T. (2000). The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an Alternative. *Journal of Educational Measurement*, *37*(1), 77–86. http://doi.org/10.1111/j.1745-3984.2000.tb01077.x

Borowsky, A., Oron-Gilad, T., Meir, A., & Parmet, Y. (2011). Drivers' perception of vulnerable road users: A hazard perception approach. *Accident Analysis & Prevention*, *44*(1), 160–166. http://doi.org/10.1016/j.aap.2010.11.029

Borowsky, A., Shinar, D., & Oron-Gilad, T. (2007). Age, skill and hazard perception in driving (Vol. 551). Presented at the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Stevenson, WA, USA.

Borowsky, A., Shinar, D., & Oron-Gilad, T. (2010). Age, skill, and hazard perception in driving. *Accident Analysis & Prevention*, *42*(4), 1240–1249. http://doi.org/10.1016/j.aap.2010.02.001

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. http://doi.org/10.1080/15389588.2011.591856

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Brown, I. D. (2002). A review of the “looked but failed to see” accident causation factor. Presented at the Behavioural Research In Road Safety: Eleventh Seminar, London, UK.

Castro, C., Padilla, J. L., Roca, J., Benítez, I., García-Fernández, P., Estévez, B., … Crundall, D. (2014). Development and Validation of the Spanish Hazard Perception Test. *Traffic Injury Prevention*, *15*(8), 817–826. http://doi.org/10.1080/15389588.2013.879125

Castro, C., Ventsislavova, P., Peña-Suarez, E., Gugliotta, A., Garcia-Fernandez, P., Eisman, E., & Crundall, D. (2016). Proactive Listening to a Training Commentary improves hazard prediction. *Safety Science*, *82*, 144–154. http://doi.org/10.1016/j.ssci.2015.09.018

Chapman, P., & Underwood, G. (1998). Visual search of driving situations: Danger and experience. *Perception*, *27*(8), 951–964. http://doi.org/10.1068/p270951

Chapman, P., Underwood, G., & Roberts, K. (2002). Visual search patterns in trained and untrained novice drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(2), 157–167. http://doi.org/10.1016/S1369-8478(02)00014-1

Chapman, P., Van Loon, E., Trawley, S., & Crundall, D. (2007). A comparison of drivers’ eye movements in filmed and simulated dangerous driving situations. *Behavioural Research in Road Safety, 2007: Seventeenth Seminar.*

Colbert, M. A. (2001). *Statistical Analysis of Multiple Choice Testing*. Air University.

Crundall, D. (2009). The Deceleration Detection Flicker Test: A measure of experience? *Ergonomics*, *52*(6), 674–684. http://doi.org/10.1080/00140130802528337

Crundall, D. (2016). Hazard prediction discriminates between novice and experienced drivers. *Accident Analysis & Prevention*, *86*, 47–58. http://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 & Prevention*, *42*(6), 2117–2124. http://doi.org/10.1016/j.aap.2010.07.001

Crundall, D., & Chapman, P. (2014, June). *The Seven Sins of Hazard Perception*. Presented at the 46th CIECA General Assembly and Congress, Dublin, Ireland. Retrieved from http://www.cieca.eu/template_events.asp?eve_id=126&lng_iso=EN

Crundall, D., Chapman, P., France, E., Underwood, G., & Phelps, N. (2005). What attracts attention during police pursuit driving? *Applied Cognitive Psychology*, *19*(4), 409–420. http://doi.org/10.1002/acp.1067

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

Crundall, D., Chapman, P., Trawley, S., Collins, L., van Loon, E., Andrews, B., & Underwood, G. (2012). Some hazards are more attractive than others: Drivers of varying experience respond differently to different types of hazard. *Accident Analysis & Prevention*, *45*, 600–609. http://doi.org/10.1016/j.aap.2011.09.049

Crundall, D., & Underwood, G. (1998). Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics*, *41*(4), 448–458. http://doi.org/10.1080/001401398186937

Crundall, D., Underwood, G., & Chapman, P. (2002). Attending to the peripheral world while driving. *Applied Cognitive Psychology*, *16*(4), 459–475. http://doi.org/10.1002/acp.806

Deery, H. A. (1999). Hazard and Risk Perception among Young Novice Drivers. *Journal of Safety Research*, *30*(4), 225–236. http://doi.org/10.1016/S0022-4375(99)00018-3

Department for Transport (2016). *Theory test: cars - GOV.UK*. Retrieved from http://www.gov.uk/theory-test/hazard-perception-test

Drummond, A. E. (2000). Paradigm lost! Paradigm gained? An Australian's perspective on the novice driver problem. Presented at the Novice Drivers Conference, Bristol, UK.

Falkmer, T., & Gregersen, N. P. (2005). A Comparison of Eye Movement Behavior of Inexperienced and Experienced Drivers in Real Traffic Environments. *Optometry & Vision Science*, *82*(8), 732–739.

Finn, P., & Bragg, B. W. E. (1986). Perception of the risk of an accident by young and older drivers. *Accident Analysis & Prevention*, *18*(4), 289–298. http://doi.org/10.1016/0001-4575(86)90043-6

Fisher, D. L., Pollatsek, A. P., & Pradhan, A. (2006). Can novice drivers be trained to scan for information that will reduce their likelihood of a crash? *Injury Prevention*, *12*(Suppl 1), i25–i29. http://doi.org/10.1136/ip.2006.012021

Galpin, A., Underwood, G., & Crundall, D. (2009). Change blindness in driving scenes. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(2), 179–185. http://doi.org/10.1016/j.trf.2008.11.002

Garay, L., Fisher, D. L., & Hancock, K. L. (2004). Effects of Driving Experience and Lighting Condition on Driving Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(19), 2290–2294.

Garay-Vega, L., & Fisher, D. L. (2005). Can Novice Drivers Recognize Foreshadowing Risks as Easily as Experienced Drivers? Presented at the Driving Assessment 2005: 3rd International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design.

Ghadiri, S. M. R., Prasetijo, J., Sadullah, A. F., Hoseinpour, M., & Sahranavard, S. (2013). Intelligent speed adaptation: Preliminary results of on-road study in Penang, Malaysia. *IATSS Research*, *36*(2), 106–114. http://doi.org/10.1016/j.iatssr.2012.08.001

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

Groeger, J. A., & Chapman, P. (1996). Judgement of Traffic Scenes: The Role of Danger and Difficulty. *Applied Cognitive Psychology*, *10*(4), 349–364. http://doi.org/10.1002/(SICI)1099-0720(199608)10:4<349::AID-ACP388>3.0.CO;2-4

Groff, L. S., & Chaparro, A. (2003). Effects of Experience and Task Relevance on the Ability to Detect Changes in a Real-World Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(13), 1605–1609. http://doi.org/10.1177/154193120304701304

Hauswald, M. (1997). Seat belt use in a developing country: covert noncompliance with a primary enforcement law in Malaysia. *Accident Analysis & Prevention*, *29*(5), 695–697. http://doi.org/10.1016/S0001-4575(97)00004-3

Horswill, M. S., Anstey, K. J., Hatherly, C. G., & Wood, J. M. (2010). The Crash Involvement of Older Drivers Is Associated with Their Hazard Perception Latencies. *Journal of the International Neuropsychological Society*, *16*(5), 939–944. http://doi.org/10.1017/S135561771000055X

Horswill, M. S., Falconer, E. K., Pachana, N. A., Wetton, M., & Hill, A. (2015). The longer-term effects of a brief hazard perception training intervention in older drivers. *Psychology and Aging*, *30*(1), 62–67. http://doi.org/10.1037/a0038671

Horswill, M. S., Hill, A., & Wetton, M. (2015). Can a video-based hazard perception test used for driver licensing predict crash involvement? *Accident Analysis & Prevention*, *82*, 213–219. http://doi.org/10.1016/j.aap.2015.05.019

Horswill, M. S., Marrington, S. A., McCullough, C. M., Wood, J., Pachana, N. A., McWilliam, J., & Raikos, M. K. (2008). The Hazard Perception Ability of Older Drivers. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *63*(4), 212–218.

Horswill, M. S., & McKenna, F. P. (2004). Drivers' hazard perception ability: situation awareness on the road. In S. Banbury & S. Tremblay, *A Cognitive approach to situation awareness: theory and application* (pp. 155–175). Aldershot, United Kingdom: Ashgate.

Horswill, M. S., Taylor, K., Newnam, S., Wetton, M., & Hill, A. (2013). Even highly experienced drivers benefit from a brief hazard perception training intervention. *Accident Analysis & Prevention*, *52*, 100–110. http://doi.org/10.1016/j.aap.2012.12.014

Huestegge, L., Skottke, E.-M., Anders, S., Müsseler, J., & Debus, G. (2010). The development of hazard perception: Dissociation of visual orientation and

hazard processing. *Transportation Research Part F: Traffic Psychology and Behaviour*, *13*(1), 1–8. http://doi.org/10.1016/j.trf.2009.09.005

Isler, R. B., Starkey, N. J., & Sheppard, P. (2011). Effects of higher-order driving skill training on young, inexperienced drivers' on-road driving performance. *Accident Analysis & Prevention*, *43*(5), 1818–1827. http://doi.org/10.1016/j.aap.2011.04.017

Isler, R. B., Starkey, N. J., & Williamson, A. R. (2009). Video-based road commentary training improves hazard perception of young drivers in a dual task. *Accident Analysis & Prevention*, *41*(3), 445–452. http://doi.org/10.1016/j.aap.2008.12.016

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. http://doi.org/10.1080/00140130802030714

Kilbey, P. (2011). *Reported Road Casualties in Great Britain: 2010 Annual Report*. UK: Department for Transport.

King, M. J., Lewis, I. M., & Abdul Hanan, S. (2011). Understanding Speeding in School Zones in Malaysia and Australia Using an Extended Theory of Planned Behaviour: The Potential Role of Mindfulness. *Journal of the Australasian College of Road Safety*, *22*(2), 56.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*(14).

Konstantopoulos, P., Chapman, P., & Crundall, D. (2010). Driver's visual attention as a function of driving experience and visibility. Using a driving simulator to explore drivers' eye movements in day, night and rain driving. *Accident*

Analysis & Prevention, *42*(3), 827–834. http://doi.org/10.1016/j.aap.2009.09.022

Lee, Y. M., Sheppard, E., & Crundall, D. (2011). [The Deceleration Detection Flicker Task applied in Malaysia]. Unpublished raw data.

Lim, P. C., Sheppard, E., & Crundall, D. (2013). Cross-cultural effects on drivers' hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, *21*, 194–206. http://doi.org/10.1016/j.trf.2013.09.016

Lim, P. C., Sheppard, E., & Crundall, D. (2014). A predictive hazard perception paradigm differentiates driving experience cross-culturally. *Transportation Research Part F: Traffic Psychology and Behaviour*, *26*, *Part A*, 210–217. http://doi.org/10.1016/j.trf.2014.07.010

Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about "weapon focus." *Law and Human Behavior*, *11*(1), 55–62. http://doi.org/10.1007/BF01044839

Lund, I. O., & Rundmo, T. (2009). Cross-cultural comparisons of traffic safety, risk perception, attitudes and behaviour. *Safety Science*, *47*(4), 547–553. http://doi.org/10.1016/j.ssci.2008.07.008

Mayhew, D. R., Simpson, H. M., & Pak, A. (2003). Changes in collision rates among novice drivers during the first months of driving. *Accident Analysis & Prevention*, *35*(5), 683–691. http://doi.org/10.1016/S0001-4575(02)00047-7

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–S23. http://doi.org/10.1016/j.jadohealth.2015.02.013

Mckenna, F., & Crick, J. (1991). Experience and expertise in hazard perception. In G. B. Grayson & J. F. Lester, *Behavioral Research in Road Safety* (pp. 39–45). Transport Research Laboratory.

McKenna, F. P., & Crick, J. (1997). *Developments in hazard perception* (No. 313). UK: Transport Research Laboratory.

McKenna, F. P., & Horswill, M. S. (1999). Hazard perception and its relevance for driver licensing. *Journal of the International Association of Traffic and Safety Sciences*, *23*(1), 36–41.

McKenna, F. P., Horswill, M. S., & Alexander, J. L. (2006). Does Anticipation Training Affect Drivers' Risk Taking? *Journal of Experimental Psychology: Applied*, *12*(1), 1–10. http://doi.org/10.1037/1076-898X.12.1.1

McKnight, A. J., & McKnight, A. S. (2003). Young novice drivers: careless or clueless? *Accident Analysis & Prevention*, *35*(6), 921–925. http://doi.org/10.1016/S0001-4575(02)00100-8

Meir, A., Borowsky, A., & Oron-Gilad, T. (2014). Formation and Evaluation of Act and Anticipate Hazard Perception Training (AAHPT) Intervention for Young Novice Drivers. *Traffic Injury Prevention*, *15*(2), 172–180. http://doi.org/10.1080/15389588.2013.802775

Mohamed, N., Mohd-Yusoff, M.-F., Othman, I., Zulkipli, Z.-H., Osman, M. R., & Voon, W. S. (2012). Fatigue-related crashes involving express buses in Malaysia: Will the proposed policy of banning the early-hour operation reduce fatigue-related crashes and benefit overall road safety? *Accident Analysis & Prevention*, *45*, *Supplement*, 45–49. http://doi.org/10.1016/j.aap.2011.09.025

Mourant, R. R., & Rockwell, T. H. (1972). Strategies of Visual Search by Novice and Experienced Drivers. *Human Factors: The Journal of the Human Factors and*

Ergonomics Society, *14*(4), 325–335. http://doi.org/10.1177/001872087201400405

Nantulya, V. M., & Reich, M. R. (2002). The neglected epidemic: road traffic injuries in developing countries. *BMJ : British Medical Journal*, *324*(7346), 1139–1141.

Nordfjærn, T., & Rundmo, T. (2009). Perceptions of traffic risk in an industrialised and a developing country. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(1), 91–98. http://doi.org/10.1016/j.trf.2008.08.003

Özkan, T., Lajunen, T., Chliaoutakis, J. E., Parker, D., & Summala, H. (2006a). Cross-cultural differences in driving behaviours: A comparison of six countries. *Transportation Research Part F: Traffic Psychology and Behaviour*, *9*(3), 227–242. http://doi.org/10.1016/j.trf.2006.01.002

Özkan, T., Lajunen, T., Chliaoutakis, J. E., Parker, D., & Summala, H. (2006b). Cross-cultural differences in driving skills: A comparison of six countries. *Accident Analysis & Prevention*, *38*(5), 1011–1018. http://doi.org/10.1016/j.aap.2006.04.006

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., & Mathers, C. (2004). *World report on road traffic injury prevention*. World Health Organization, Geneva.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. http://doi.org/10.3389/neuro.11.010.2008

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. http://doi.org/10.1163/156856897X00366

Pelz, D. C., & Krupat, E. (1974). Caution profile and driving record of undergraduate males. *Accident Analysis & Prevention*, *6*(1), 45–58. http://doi.org/10.1016/0001-4575(74)90015-3

Pollatsek, A., Narayanaan, V., Pradhan, A., & Fisher, D. L. (2006). Using Eye Movements to Evaluate a PC-Based Risk Awareness and Perception Training Program on a Driving Simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 447–464. http://doi.org/10.1518/001872006778606787

Poulsen, A. A., Horswill, M. S., Wetton, M. A., Hill, A., & Lim, S. M. (2010). A Brief Office-Based Hazard Perception Intervention for Drivers With ADHD Symptoms. *Australian and New Zealand Journal of Psychiatry*, *44*(6), 528–534. http://doi.org/10.3109/00048671003596048

Queensland Treasury and Trade. (2012). *Population Growth Highlights and Trends, Queensland 2012*.

Quimby, A. R., Maycock, G., Carter, I. D., Dixon, R., & Wall, J. G. (1986). *Perceptual abilities of accident involved drivers* (Research Report No. 27). Crowthorne, United Kingdom: Transport Research Laboratory.

Rohayu, S., Sharifah Allyana, S. M. R., Jamilah, M. M., & Wong, S. V. (2012). *Predicting Malaysian Road Fatalities for Year 2020* (MRR No. 06/2012). Malaysian Institute of Road Safety Research (MIROS).

Sagberg, F., & Bjørnskau, T. (2006). Hazard perception and driving experience among novice drivers. *Accident Analysis & Prevention*, *38*(2), 407–414. http://doi.org/10.1016/j.aap.2005.10.014

Scialfa, C. T., Borkenhagen, D., Lyon, J., & Deschênes, M. (2013). A comparison of static and dynamic hazard perception tests. *Accident Analysis & Prevention*, *51*, 268–273. http://doi.org/10.1016/j.aap.2012.12.006

Scialfa, C. T., Borkenhagen, D., Lyon, J., Deschênes, M., Horswill, M. S., & Wetton, M. A. (2012). The effects of driving experience on responses to a static hazard perception test. *Accident Analysis & Prevention*, *45*, 547–553. http://doi.org/10.1016/j.aap.2011.09.005

Scialfa, C. T., Deschênes, M. C., Ference, J., Boone, J., Horswill, M. S., & Wetton, M. A. (2011). A hazard perception test for novice drivers. *Accident Analysis & Prevention*, *43*(1), 204–208. http://doi.org/10.1016/j.aap.2010.08.010

Sivak, M., Soler, J., Tränkle, U., & Spagnhol, J. M. (1989). Cross-cultural differences in driver risk-perception. *Accident Analysis & Prevention*, *21*(4), 355–362. http://doi.org/10.1016/0001-4575(89)90026-2

Soliday, S. M. (1974). Relationship between age and hazard perception in automobile drivers. *Perceptual and Motor Skills*, *39*(1), 335–338.

Sullman, M. J. M., Stephens, A. N., & Yong, M. (2014). Driving anger in Malaysia. *Accident Analysis & Prevention*, *71*, 1–9. http://doi.org/10.1016/j.aap.2014.04.019

Sullman, M. J. M., Stephens, A. N., & Yong, M. (2015). Anger, aggression and road rage behaviour in Malaysian drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, *29*, 70–82. http://doi.org/10.1016/j.trf.2015.01.006

Toroyan, T. (2009). *Global Status Report on Road Safety: Time for Action*. Geneva: World Health Organization.

Toroyan, T. (2013). *Global status report on road safety 2013: supporting a decade of action*. Geneva: World Health Organization.

Underwood, G., Chapman, P., Berger, Z., & Crundall, D. (2003). Driving experience, attentional focusing, and the recall of recently inspected events. *Transportation Research Part F: Traffic Psychology and Behaviour*, *6*(4), 289–304. http://doi.org/10.1016/j.trf.2003.09.002

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. http://doi.org/10.1016/S1369-8478(02)00008-6

Underwood, G., Chapman, P., Brocklehurst, N., Underwood, J., & Crundall, D. (2003). Visual attention while driving: sequences of eye fixations made by experienced and novice drivers. *Ergonomics*, *46*(6), 629–646. http://doi.org/10.1080/0014013031000090116

Underwood, G., Crundall, D., & Chapman, P. (2011). Driving simulator validation with hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(6), 435–446. http://doi.org/10.1016/j.trf.2011.04.008

Underwood, G., Ngai, A., & Underwood, J. (2013). Driving experience and situation awareness in hazard detection. *Safety Science*, *56*, 29–35. http://doi.org/10.1016/j.ssci.2012.05.025

Underwood, G., Phelps, N., Wright, C., Van Loon, E., & Galpin, A. (2005). Eye fixation scanpaths of younger and older drivers in a hazard perception task. *Ophthalmic and Physiological Optics*, *25*(4), 346–356. http://doi.org/10.1111/j.1475-1313.2005.00290.x

Vlakveld, W. P. (2014). A comparative study of two desktop hazard perception tasks suitable for mass testing in which scores are not based on response latencies.

Transportation Research Part F: Traffic Psychology and Behaviour, *22*, 218–231. http://doi.org/10.1016/j.trf.2013.12.013

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 & Prevention*, *39*(6), 1177–1185. http://doi.org/10.1016/j.aap.2007.03.003

Watts, G. R., & Quimby, A. R. (1979). *Design and Validation of a Driving Simulator for use in Perceptual Studies* (No. 907). Transport Research Laboratory.

Wells, P., Tong, S., Sexton, B., Grayson, G., & Jones, E. (2008). *Cohort II: a study of learner and new drivers: volume 1* (Road Safety Research Report No. 81). London: Department for Transport.

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 & Prevention*, *43*(5), 1759–1770. http://doi.org/10.1016/j.aap.2011.04.007

Wetton, M. A., Hill, A., & Horswill, M. S. (2013). Are what happens next exercises and self-generated commentaries useful additions to hazard perception training for novice drivers? *Accident Analysis & Prevention*, *54*, 57–66. http://doi.org/10.1016/j.aap.2013.02.013

Wetton, M. A., Horswill, M. S., Hatherly, C., Wood, J. M., Pachana, N. A., & Anstey, K. J. (2010). The development and validation of two complementary measures of drivers' hazard perception ability. *Accident Analysis & Prevention*, *42*(4), 1232–1239. http://doi.org/10.1016/j.aap.2010.01.017

Yeung, J. S., & Wong, Y. D. (2015). Effects of driver age and experience in abrupt-onset hazards. *Accident Analysis & Prevention*, *78*, 110–117. http://doi.org/10.1016/j.aap.2015.02.024

Yusoff, M. F. M., Baki, M. M., Mohamed, N., Mohamed, A. S., Yunus, M. R. M., Ami, M., … Ishak, A. I. (2010). Obstructive Sleep Apnea Among Express Bus Drivers in Malaysia: Important Indicators for Screening. *Traffic Injury Prevention*, *11*(6), 594–599. http://doi.org/10.1080/15389588.2010.505255